Model-Based Recursive Partitioning for Stratified and Personalised Medicine

---

Dissertation

zur
Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der
Mathematisch-naturwissenschaftlichen Fakultät
der
Universität Zürich

von

Heidi Seibold

aus
Deutschland

Promotionskommission

Prof. Dr. Torsten Hothorn (Vorsitz)
Prof. Dr. Achim Zeileis
Prof. Dr. Milo Puhan

Zürich, 2018

Zusammenfassung

In klinischen Studien werden Patienten zufällig in (üblicherweise zwei) Behandlungsgruppen eingeteilt. Dabei ist eine der Behandlungen in der Regel eine Neue, bei der man die Wirksamkeit überprüfen möchte. Die andere Behandlung is meist ein Behandlungsstandard für die vorliegende Krankheit oder ein Placebo. Nach Studienende werden die erhobenen Daten über den Krankheitsverlauf seit Behandlungsbeginn analysiert, um zu klären ob und inwieweit die neue Behandlungsmethode besser ist. Klassische statistische Methoden für die Analyse von randomisierten klinischen Studien nehmen an, dass der Behandlungseffekt konstant, also für alle Patienten gleich, ist und – was noch wichtiger ist –, dass der Behandlungseffekt auch für alle zukünftigen Patienten gleich sein wird. Diese Annahme ist besonders bei komplexen Erkrankungen und heterogenen Patientenpopulationen unrealistisch. Neue statistische Methoden werden benötigt, um herauszufinden, ob sich Behandlungseffekte zwischen Patienten unterscheiden und, wenn dies der Fall ist, welche Patientencharakteristiken den Behandlungseffekt beeinflussen.

Für die Auffindung von Subgruppen, die sich hinsichtlich des Behandlungseffekts unterscheiden und bei denen innerhalb der Subgruppe der Bahandlungseffekt gleich ist, schlagen wir modellbasierte Bäume als geeignete statistische Methode vor. Subgruppenanalysen werden im Folgenden auch als *stratifizierte Medizin* bezeichnet. Für *personalisierte Medizin*, also die Schätzung von personalisierten Behandlungseffekten, können modellbasierte Zufallswälder genutzt werden. Die Zufallswälder liefern ein Ähnlichkeitsmass, das bestimmt, wie ähnlich sich Patienten hinsichtlich des Behandlungseffekts sind. Dieses Ähnlichkeitsmass wird dann genutzt, um in Kombination mit Modellen den Behandlungseffekt zu schätzen.

Manchmal ist bereits vor Studienbeginn bekannt, dass bestimmte Patientencharakteristiken die Zielgrösse direkt beeinflussen und zwar unabhänging von der Behandlung. In diesen Situationen ist es von Interesse, den prognostischen Effekt dieser Patientencharakteristiken gobal zu schätzen und nicht pro Subgruppe, wie in den oben genannten modellbasierten Bäumen. Für solche Anwendungen empfehlen wir Bäume mit partiell additiven (generalisierten) linearen Modellen, auch PALM trees – partially additive (generalised) linear model trees – genannt. In diesen Bäumen werden manche Modellparameter pro Subgruppe geschätzt (in der Regel Interzept und Behandlungseffekt) und andere global (Effekte von bekannten prognostischen Patientencharakteristiken).

Die PRO-ACT Datenbank enthält Daten von mehreren klinischen Studien zu Amyotropher Lateralsklerose (ALS), eine komplexe neurodegenerative Krankeit für die auf dem Markt nur ein Medikament zugelassen ist. Für dieses Medikament, genannt Riluzol oder auch Rilutek, konnte nur einen moderate Verlängerung der Lebenserwartung nachgewiesen werden und es ist unklar, ob alle Patienten von der Einnahme des Medikaments profitieren. Wir analysierten die PRO-ACT Daten mit modellbasierten Bäumen und Zufallswäldern und fanden Anzeichen dafür,

dass sich die Behandlungseffekte zwischen Patienten unterscheiden. Die Analysen werden im Folgeneden gezeigt, um die Methoden zu veranschaulichen.

Für alle erwähnten Methoden ist Software verfügbar. Modellbasierte Bäume und Zufallswälder sind im R Paket `partykit` implementiert. Die Zielgruppe für das `partykit`-Paket ist breit und daher nicht bestmöglich auf die Anwendung für stratifizierte und personalisierte Medizin abgestimmt. Daher bieten wir auch ein Zusatzpaket mit dem Namen `model4you` an, das leicht zu bedienen ist und interpretierbare Graphiken erzeugt. PALM trees sind im `palmtree`-Paket implementiert.

Abstract

In randomised clinical trials patients are randomly assigned to a new treatment of interest or a control, e.g. the standard of care or a placebo. At the end of the study, data about the course of the disease of the patients is analysed to answer the question whether the new treatment is better than the control. Established statistical procedures for the analysis of primary endpoints in randomised clinical trials assume that there is a universal, i.e. constant, treatment effect that applies to all patients in the trial and - even more importantly - to all future patients potentially to be treated with the new treatment under consideration. For complex diseases or heterogeneous patient populations this assumption may be incorrect and novel statistical methods are needed to discover if treatment effects differ across patients, and if so, which patient characteristics influence treatment effects.

We propose model-based trees as a method for *stratified medicine* – i.e. for the discovery of patient subgroups, where within subgroups the treatment effects are the same and between subgroups treatment effects differ. Using ensembles of model-based trees (model-based forests) we can detect similarities between patients in terms of treatment effects and use the similarity measure to estimate personalised treatment effects (*personalised medicine*).

Sometimes patient characteristics are known to have a direct effect on the primary outcome, irrespective of treatment. In these situations it may be relevant to estimate global effects for these prognostic factors. As this is not possible in regular model-based trees, we propose partially additive (generalised) linear model trees (PALM trees) as a variation of classical model-based trees where some effect estimates are stratified (usually intercept and treatment effect) and some are global (effects of known prognostic factors).

The PRO-ACT database contains data from several clinical trials about amyotrophic lateral sclerosis (ALS), which is a complex neurodegenerative disease for which only one drug, Riluzole, is available on the market. The drug has been shown to be only moderately effective and it is unclear if all patients benefit from it. We analysed the PRO-ACT data using model-based trees and forests and found evidence that treatment effects vary across patients. We use the analysis of the PRO-ACT data to guide the reader through the different methods.

Open source software for all proposed methods is available. The `partykit` package implements basic model-based trees and random forests in R. The functionality in the `partykit` package is aimed at a broad audience. To make it easy for users interested in stratified and personalised medicine, we offer an add-on package called `model4you`, which focuses on user friendliness and interpretability of the results. PALM trees are implemented in package `palmtree`.

Thesis outline

Preface

This thesis is submitted under the Ph.D. program in "Epidemiology and Biostatistcis" at the University of Zurich. The Ph.D. project was funded by the Swiss National Science Foundation (Grant 205321_163456) and a supporting mobility grant (205321_163456/2) for a six month research stay at the University of Innsbruck. Four articles originating from the project are bundled in this thesis. Open-source software is available in the form of R packages for all methods developed.

I would like to thank all those who have supported me and my research, in particular ...

... my supervisors and mentors Torsten Hothorn and Achim Zeileis, who helped me become the researcher I am now and made this Ph.D. such a great experience.

... Andrea Farnham and Isaac Gravestock, who are not only great colleagues helping me with english corrections and computer problems, but have become close friends with whom I enjoyed my time outside the office and in the mountains.

... the entire Ph.D. program for being such a great group of people.

... Christoph Molnar, who has been at my side as partner, friend and nerd.

... My parents Gerda and Albert Seibold, who raised me to become a strong woman without fear of maths and computers.

Zürich, December 2017 Heidi Seibold

Introduction

This thesis covers model-based trees and forests for the estimation of stratified and personalised treatment effects. It explains why model-based recursive partitioning – although not exclusively designed for it – is so useful for stratified and personalised medicine and describes how it works.

This introduction unfolds as follows: Section 1 defines the terms *stratified* and *personalised medicine* and documents the importance of the topic and respective statistical methods. Section 2 introduces relevant statistical methods. Parametric models are discussed first (see Section 2.1) as they have traditionally been used for the estimation of treatment effects and form the basis of model-based recursive partitioning. Then Section 2.2 gives an introduction into model-based recursive partitioning including model-based trees, PALM trees and model-based forests. The usage is illustrated on data from ALS patients. Section 3 gives a brief overview of the four papers forming the output of this Ph.D. project. My work has already sparked some interest in the academic as well as in the pharmaceutical community, which is documented in Section 4.

1 Stratified and personalised medicine

Personalised medicine (also called *precision medicine*) has been a buzzword in recent years, especially following Barack Obama's 2015 speech introducing the US *preci-*

sion medicine initiative[1] , where he said:

> *"And that's why we're here today. Because something called precision medicine – in some cases, people call it personalized medicine – gives us one of the greatest opportunities for new medical breakthroughs that we have ever seen. Doctors have always recognized that every patient is unique, and doctors have always tried to tailor their treatments as best they can to individuals. You can match a blood transfusion to a blood type. That was an important discovery. What if matching a cancer cure to our genetic code was just as easy, just as standard? What if figuring out the right dose of medicine was as simple as taking our temperature?"*

Obama speaks about "delivering the right treatments [...] to the right person", but at the same time he speaks about accurately diagnosing diseases. This shows that the term *personalised medicine* is not clearly defined (see also Schleidgen et al., 2013). In the context of genomics or rare diseases it can be understood as the accurate identification of a disease. In other contexts it is understood as the identification of the optimal treatment for each patient or, as in this thesis, the estimation of personalised treatment effects. Note that patient characteristics influencing the treatment effect might as well signal that patients are suffering from different sub-diseases. For example, a gene mutation could lead to a certain sub-disease that should be treated differently. Identifying the sub-disease may then be a side product of estimating personalised treatment effects, but is not the original aim.

The terms *personalised medicine* and *stratified medicine* are often used synonymously. In this work I differentiate between stratified medicine, where subgroup-wise treatment effects are estimated and the focus lies on identifying the patient characteristics defining the subgroups, and personalised medicine, where we go a step further and estimate a treatment effect for each individual. Treatment effects may well be homogeneous among a study population and statistical methods for personalised and stratified medicine must not only be able to identify patient characterstics interacting with the treatment but must also be able to identify when there is no heterogeneity.

The aim of my work is to detect patient characteristics that influence the treatment effect. In the medical literature these patient characteristics are commonly referred to as predictive factors (see e.g. Italiano, 2011). Patient characteristics affecting the outcome or progression, called *prognostic factors*, are of secondary interest, but are also retrieved in model-based recursive partitioning, giving a holistic view on health of patients. For a visual reminder of the difference between predictive and prognostic factor see Figure 1.

[1] https://obamawhitehouse.archives.gov/the-press-office/2015/01/30/remarks-president-precision-medicine

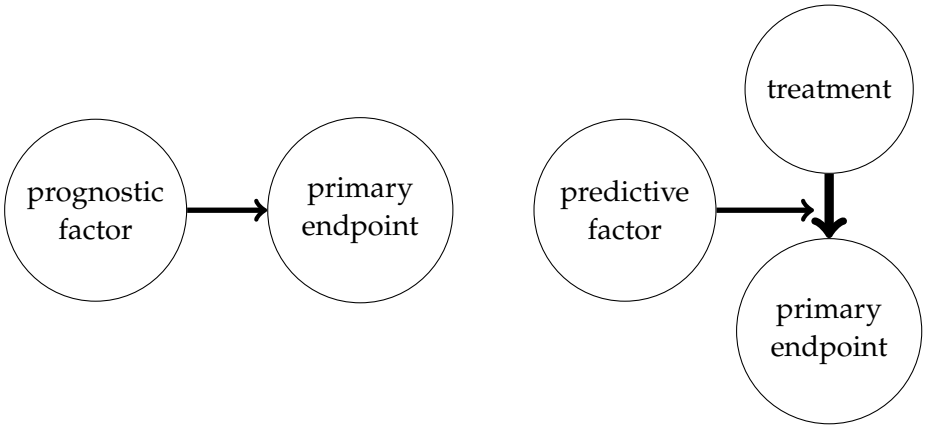


Figure 1: Definition of prognostic and predictive factors.

The regulatory agencies EMA and FDA are recognising that stratified and personalised medicine is an important topic for drug development (see reports US Food and Drug Administration, 2013; European Medicines Agency, 2014) and the topic is currently frequently discussed in academia (e.g. Weisberg, 2015), pharmaceutical industry (Das, 2017) and even the media (e.g. New York Times Editorial Board, 2015). The standard way of analysing clinical trials, however, is estimating average treatment effects, i.e. one treatment effect per study population. Since their rise in popularity after the first randomised controlled trial conducted in 1946, randomised clinical trials have helped medical researchers to develop and evaluate new treatments for many diseases. Hence randomised clinical trials and average treatment effects are rightfully popular. In recent years, drug and treatment development has become more difficult, since treatments for less complex diseases are already known (Weisberg, 2015). Developing treatments for complex diseases with unknown underlying biological processes or unknown subcategories is extremely difficult. One treatment may work for a patient with disease subcategory A but not for a patient with disease subcategory B. The finding that trastuzumab should only be used for breast cancer patients with HER-2 protein overexpression (Slamon et al., 2001; Frueh and Gurwitz, 2004) is a famous example. For diseases where patients react to the same treatment in different ways, medical research needs statistical methods that can detect and estimate heterogeneous treatment effects. These exploratory methods can then influence hypothesis generation and thus aid planning of more specific and informed clinical trials.

The articles in this thesis each deal with stratified and/or personalised medicine and focus on statistical methods that detect patient characteristics influencing the treatment effect and then estimate stratified or personalised treatment effects, so that future clinical trials can be designed using well informed hypotheses, generated with the help of statistically sound data driven methods.

2 Statistical methods for stratified and personalised medicine

This section describes statistical methods used to analyse clinical trials. Parametric and semi-parametric models are the standard for estimating average treatment effects and these same models can be used to partition the data using model-based recursive partitioning to retrieve stratified or personalised treatment effects.

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease that leads to shrinking of muscles. Patients suffer from speaking and swallowing problems and, ultimately, breathing impairment. ALS patients have a low survival expectancy and the only approved drug against ALS – Riluzole – has been shown to prolong life expectancy by merely 2 months (European Medicines Agency, 2012). This gives a first impression on how complex this disease is. Two of the articles in this thesis analyse data from ALS patients. The data are a collection of clinical trials about ALS (Pooled Resource Open–Access Clinical Trials database, Atassi et al., 2014) and next to treatment information also contain information about baseline patient characteristics. The question is, which patient characteristics lead to differences in Riluzole treatment effects and what are the treatment effects for given patients. In the following I will demonstrate the methods by analysing the Riluzole effect on the survival of ALS patients.

2.1 Models for the estimation of treatment effects

Parametric models such as linear models, generalised linear models or accelerated failure time models and the semiparametric Cox model are the most commonly used models for the estimation of treatment effects in clinical trials. The model,

$$\mathcal{M}((Y, \mathbf{X}), \boldsymbol{\vartheta}), \tag{1}$$

to be used for this **primary analysis** is generally defined in the study protocol including the primary endpoint Y, covariates $\mathbf{X}$ and parameters $\boldsymbol{\vartheta}$. The covariates in this model always contain the treatment indicator

$$X_{\text{treatment}} = \begin{cases} 1 & \text{if patient receives the new treatment} \\ 0 & \text{else.} \end{cases} \tag{2}$$

Possibly further relevant covariates can be included. In the simplest case the parameters $\boldsymbol{\vartheta}$ include an intercept (or baseline hazard) and a treatment effect. To

	estimate	2.5 %	97.5 %
α_1	6.7070	6.6438	6.7702
β	0.1073	0.0314	0.1832
$\log(\alpha_2)$	−0.5833	−0.6365	−0.5302

Table 1: Parameter estimates including confidence intervals of overall Riluzole effect Weibull model in the PRO-ACT data.

estimate the parameters, the objective function Ψ – e.g. the negative log-likelihood – is minimised or equivalently the score equation

$$\sum_{i=1}^{n} \psi((y, \mathbf{x})_i, \boldsymbol{\vartheta}) = 0 \tag{3}$$

is solved. The score contributions $\psi((y, \mathbf{x})_i, \boldsymbol{\vartheta})$ – with $i = 1, \ldots, n$ and $n =$ number of patients in the trial – are the derivatives of the contributions to the objective function (e.g. log-likelihood contributions).

To estimate the Riluzole effect on the survival of ALS patients, we use a Weibull model. This means, that the parameter vector $\boldsymbol{\vartheta}$ contains three parameters: two parameters defining the baseline hazard, which we call α_1 (intercept) and α_2 (scale parameter), and the treatment effect parameter, β. Weibull models are estimated via maximum-likelihood estimation, wich means that our objective function is the negative log-likelihood. The estimated parameters of this model are shown in Table 1 and the estimated survivor curves for the two treatment groups are shown in Figure 2. Both suggest that Riluzole leads to a slightly prolonged survival. The estimated difference in median survival between the two treatment groups is 75.5 days (i.e. about 2.5 months).

Secondary analyses estimating the treatment effect in different patient subgroups are usually conducted using the same model but using different subsets of the data and these subsets are commonly decided upon by experts. If, for example, the experts think that the treatment may work differently in old and young patients, they decide on a certain age defining the cutpoint where everyone below this age is assigned to the young patient group and vice versa. This procedure may be problematic for several reasons: (1) The expert might not know about certain patient characteristics interacting with the treatment effect; (2) higher order interactions might play a role, i.e. interactions between more than one patient characteristic and the treatment; (3) there might be a better cutpoint than the one defined by the expert or the interaction might be a smooth function instead of a step function with a certain cutpoint. All of these problems can be addressed

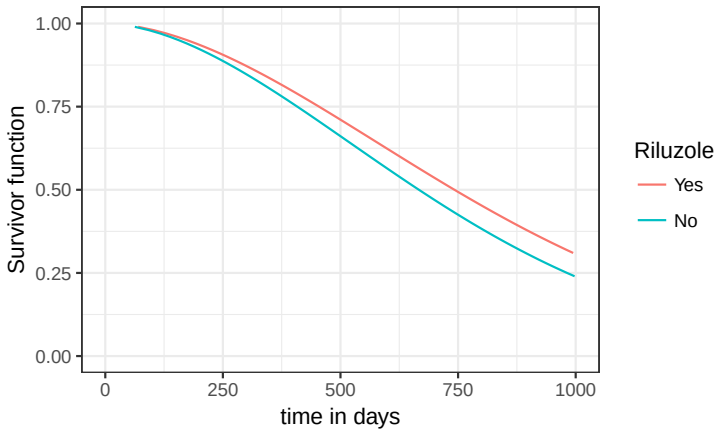


Figure 2: Survivor functions of ALS patients in the two treatment groups estimated by a Weibull model usign the PRO-ACT data.

by using a data driven method such as model-based trees and forests instead of expert knowledge.

2.2 Model-based recursive partitioning for the estimation of stratified and personalised treatment effects

Trees are designed to detect even high order interactions in a data driven way. This is why, even though they are not classically used for the analysis of clinical trials, they have been proposed by various authors as good methods for stratified medicine (see e.g. Dusseldorp and Van Mechelen, 2013; Foster et al., 2011; Negassa et al., 2005). Trees recursively split observations into subgroups where observations within a subgroup are similar and between subgroups different. Usually the similarity is defined in terms of the response. The splits are implemented based on split variables, which, in this work, are the patient characteristics. Trees can be visualised in an easy to understand and interpretable fashion. A random forest consists of an ensemble of trees. The trees in the forest differ from each other, because each tree is based on a subsample or bootstrap sample of the original data and for each split only a subset of split variables is made available. The trees in the forest have an even higher variablity since they are usually grown larger (deeper) than single trees. By introducing variation in the single trees and then averaging or synthesising the results, predictive performance of random forests is often better that of single trees (Breiman, 2001; Hastie et al., 2009). However, in comparison to trees, random forest are not interpretable.

Classical trees and random forests do not allow the user to focus on interactions with the treatment. For this, a combination of models and trees as in model-based

recursive partitioning is useful. This combination provides the best of both the tree world and the model world: Having models that are accepted and understood in the medical community and at the same time being able to find treatment $\times$ subgroup interactions in a data driven way.

The general idea of model-based recursive partitioning is to partition the data based on instabilities in the model parameters. In the application for stratified and personalised medicine, we are especially interested in instabilities in the treatment effect, but commonly model-based recursive partitioning also finds instabilities in the intercept. The first leads to the detection of predictive factors, the latter to the detection of prognostic factors. The first step of the algorithm is to compute the model for all n patients of the study sample as in Equation (3). Parameters of this model are considered unstable, if the corresponding scores (partial derivative of the contributions to the objective function) do not fluctuate randomly around zero but are correlated to at least one patient characteristic. The intuition for this is that the scores are residuals and we aim to estimate the optimal model. Scores that are correlated with patient characteristics indicate that certain information was not taken into account. The lowest p-value in tests of independence between the scores and each patient characteristic determines the patient characteristic in which to implement the split. The actual split point can either be found using again an independence test or by maximising the sum of objective functions of the models in the resulting subgroups. The algorithm recursively proceeds with the model estimation, testing and splitting in each subgroup. It stops when either no p-value is smaller than a predefined significance level (Bonferroni correction can be used here to adjust for multiple testing) or another stopping criterion – such as minimum number of observations in the subgroups – is fulfilled.

Model-based trees use the above algorithm and estimate one model per subgroup to obtain the stratified treatment effects. This is equivalent to solving the weighted score equation,

$$\sum_{i=1}^{n} w_{ig} \psi((y, \mathbf{x})_i, \boldsymbol{\vartheta}_g) = 0, \tag{4}$$

$$\text{with } w_{ig} = \begin{cases} 1 & \text{if patient } i \text{ is in subgroup } g \\ 0 & \text{else.} \end{cases}$$

By estimating the parameters per subgroup, where subgroups are defined by combinations of (binary) rules based on patient characteristics, we estimate parameter $\times$ patient characteristics interactions that are step functions depending on patient

characteristics, i.e.

$$\boldsymbol{\vartheta}(\mathbf{z}_i) = \begin{cases} \boldsymbol{\vartheta}_1 & \text{if patient } i \text{ with patient characteristics } \mathbf{z}_i \text{ is in subgroup 1} \\ \quad \vdots & \\ \boldsymbol{\vartheta}_g & \text{if patient } i \text{ with patient characteristics } \mathbf{z}_i \text{ is in subgroup } g. \end{cases} \tag{5}$$

By stopping based on (Bonferroni corrected) p-values and controlling the probability of incorrectly rejecting the null hypothesis of no parameter instability, the algorithm ensures that splits are rarely implemented, if the the true model parameters are the same for all patients. In this case, all patients are in the same "subgroup" and Equation (4) equals Equation (3). Model-based trees work well when clear treatment $\times$ subgroup interactions exist and effects of potential unknown prognostic factors can be approximated well by step functions. If prognostic factors are known, they can be included in the model. In this case a variation of model-based trees called partially additive linear model (PALM) trees can be used. **PALM trees** allow for prognostic factors to be included in the model in a way that their effect estimates are the same across all subgroups. So instead of a linear predictor $\mathbf{x}_i^\top \boldsymbol{\vartheta}_g$ with all parameters depending on the tree structure, the linear predictor is

$$\mathbf{x}_i^\top \begin{pmatrix} \boldsymbol{\theta}_g \\ \boldsymbol{\gamma} \end{pmatrix} \tag{6}$$

with parameters $\boldsymbol{\theta}_g$ (e.g. intercept and treatment effect) depending on the tree structure and parameters $\boldsymbol{\gamma}$ (effects of known prognostic factors) being fixed across all patients.

Figure 3 shows the model-based tree for the ALS data. Note that for this tree one of the stopping criteria was to have maximally four subgroups. The figure shows the split variables forming the subgroups, including the p-value of the independence test in the ovals. The numbers on the lines specify the split point. The tree defines four subgroups: patients who are of age 43 or younger, patients who are between 43 and 55.7 years old, patients who are older than 55.7 and for whom the time between disease onset and treatment start is 757 days or less, and patients who are older than 55.7 and for whom the time between disease onset and treatment start is more than 757 days. For each subgroup the estimated parameters and "confidence intervals", the number of observations (n) and the survivor curves in both treatment groups are given. The interpretation of confidence intervals here is unclear due to the variable and split point selection which is done prior to model estimation (for a discussion of the problem see Leeb and Pötscher, 2005). We suggest to not use them for inference but as a measure of variability. The subgroup with patients of age 43 or younger shows very similar survivor curves for both treatment groups, which suggests that Riluzole has no effect on the survival of young patients. A similar picture is shown for the subgroup of patients who are

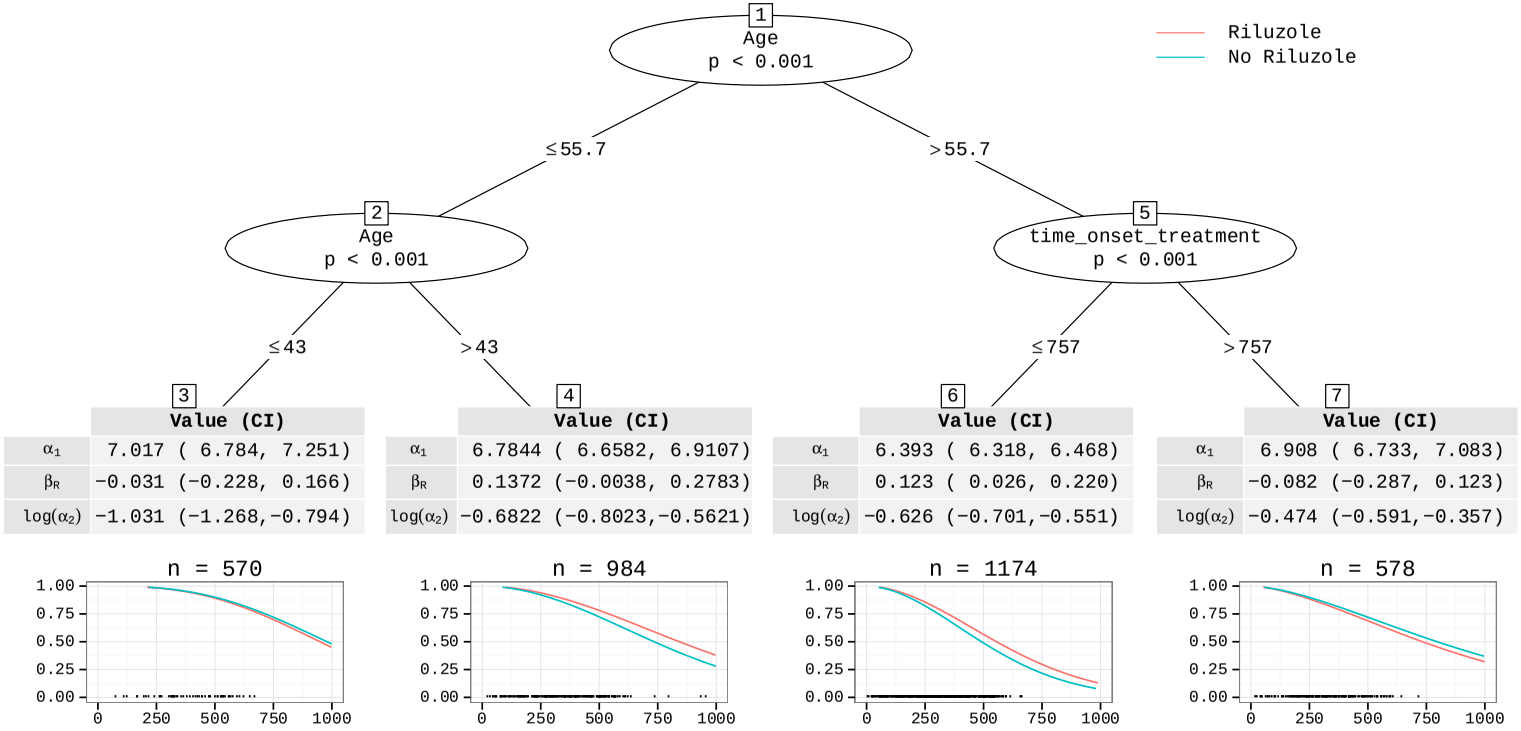


Figure 3: Model-based tree for ALS data Weibull model.

over 55.7 years old and for whom the time between disease onset and treatment start was longer than 757 days. The other two subgroups show a higher treatment effect than the overall treatment effect (see Table 1).

Patients with slow disease progression were mostly included late in the studies, which means that the time between disease onset and treatment start is a surrogate for the speed of disease progression and is thus a known prognostic factor also for the survival of patients. In this case it could be a good idea to include the time between disease onset and treatment start as a covariate in the model and fixing it across subgroups, which corresponds to the idea of PALM trees (the current implementation of the PALM tree algorithm is for generalised linear models and linear models only).

Model-based forests, like classical random forests, compute an ensemble of relatively deep trees. The usage of model-based forests, however, differs from the classical forests: Model-based forests are used to estimate the similarity of patients in terms of model parameters. In model-based trees, patients are assigned to the same subgroup if they have the same (or similar) model parameters. In this sense, the similarity between patients is one if they are in the same subgroup and zero if they are in different subgroups. A weighted model with the binary similarity measure as model weights is estimated for each subgroup g as shown in Equation (4). With model-based forests we can obtain a similarity measure that ranges between zero and the number of trees. The similarity between two patients is then the number of times they are assigned to the same subgroup in the given trees. In accordance with the stratified models, the personalised model for a new patient k is estimated by solving

$$\sum_{i=1}^{n} w_{ik} \psi((y, \mathbf{x})_i, \boldsymbol{\vartheta}_k) = 0 \tag{7}$$

with w_{ik} = number of times patients i and k are in the same subgroup. If patient k is not a *new* patient but was in the data set used to compute the forest, the weights should be computed out-of-bag to avoid including information from patient k in the model of patient k. Using the similarity weights obtained by model-based forests allow for flexible estimation of personalised models. This can lead to complex treatment $\times$ patient characteristics interactions – and more general a complex form of $\boldsymbol{\vartheta}(\mathbf{z})$ – in contrast to model-based trees where the interactions are always step functions (see Equation (5)). Since trees in random forests are grown larger than in single trees, we need to implement a possibility to detect if the personalised models are better than the base model (a single model estimating the average treatment effect). The difference in objective functions from the personalised models and the global model (estimated as in Equation (3)) gives an

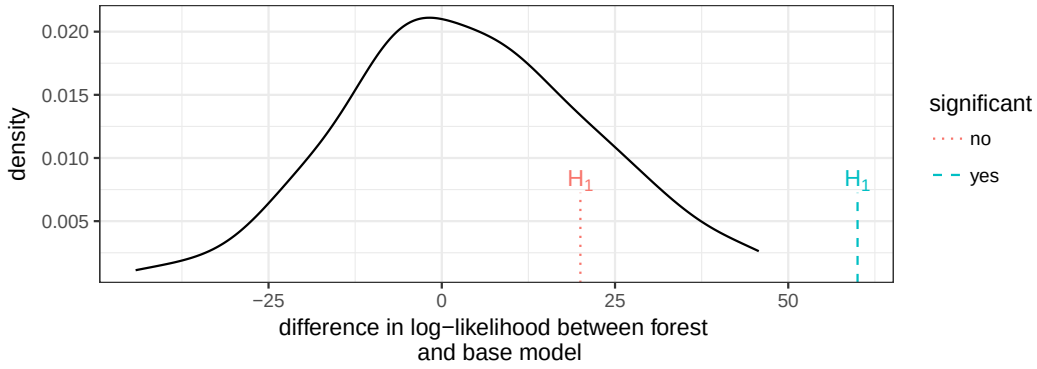


Figure 4: Illustration: test whether personalised models improve upon base model. If the true difference is high in comparison to the null distribution (density curve shown), we assume that the personalised models fit the data better than the base model.

idea of improvement. Using parametric bootstrap samples from the global model and computing again the difference in objective functions gives the distribution of differences under the null hypothesis "the base model is the correct model" (see density curve in Figure 4 as an example), which allows us to perform a significance test. If the difference in the original data is high in comparison to the differences under the null hypothesis (e.g. value corresponding to the blue line in Figure 4) the personalised models are an improvement in comparison to the base model.

For the ALS data the test suggests that the personalised models are better than the base model shown in Table 1 and Figure 2. Figure 5 shows the distribution under the null hypothesis of no difference in log-likelihood between the base model and the personalised models on the left hand side. On the right hand side the actual difference is added (denoted by H_1). The difference in log-likelihoods between the personalised models obtained trough the forest and the base model is much larger than any of the differences under the null hypothesis. Figure 6 shows the dependence plots for the personalised models computed. Different patient characteristics are plotted on the x-axis. The difference in estimated median survival between taking or not taking Riluzole for every patient is plotted on the y-axis. In normal linear models the personalised treatment effect could be plotted directly on the y-axis, as it can be interpreted as the expected improvement (or worsening) by taking the treatment instead of the control. For the Weibull model the treatment effect is the hazard ratio, which is hard to interpret – especially in comparison to other hazard ratios with different baseline hazards – which is why we chose the difference in median survival instead. The difference in median survival increases with increasing age up to 55 years and then goes down and flattens. Weak patients have a greater variability in treatment efficacy than patients who are not weak at

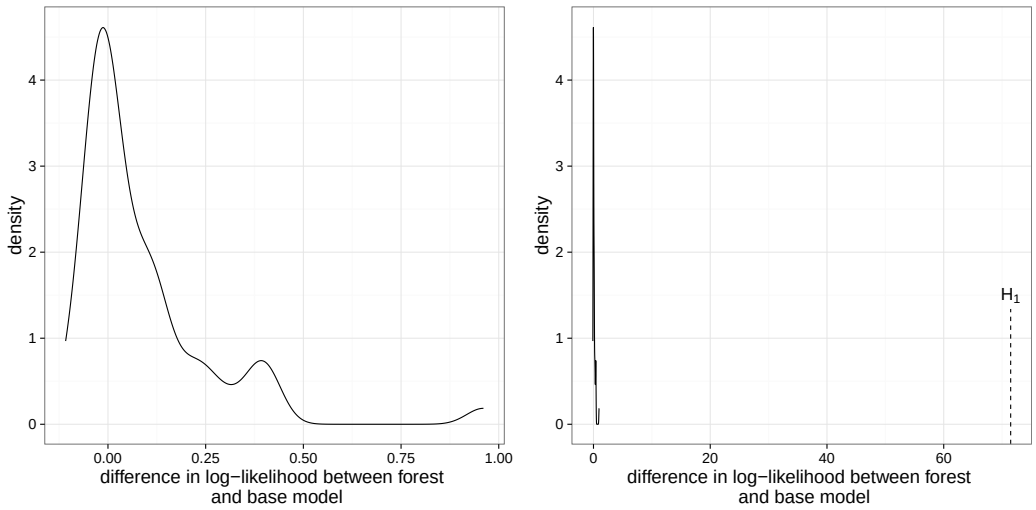


Figure 5: Test whether personalised Weibull models improve upon base model for the PRO-ACT database. Density curve shows the distribution under the null hypothesis. The two plots show the same on different scales (see changes in x-axis).

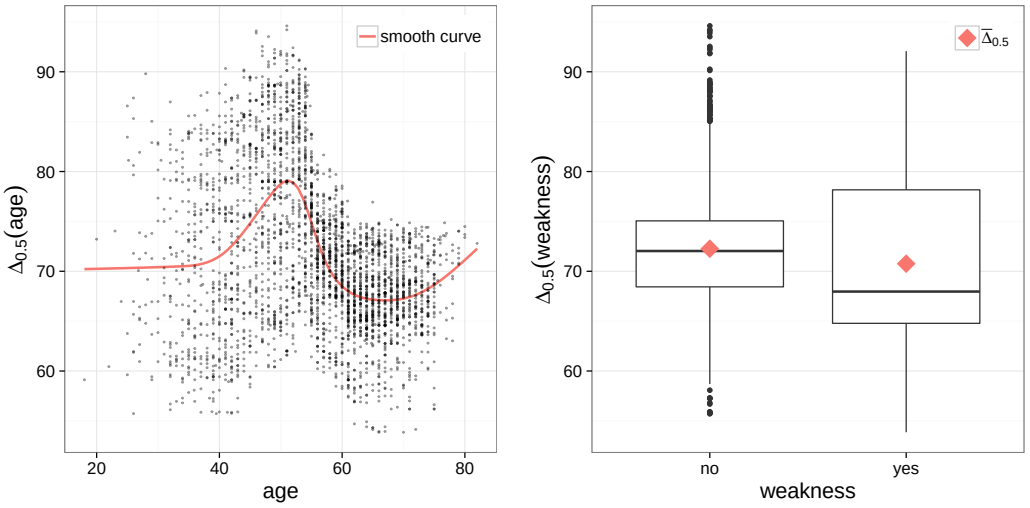


Figure 6: Dependence plots for for age and weakness obtained from personlised Weibull models for ALS patients in the PRO-ACT database.

baseline.

Summary: In this section we discussed conventional methods for the analysis of randomised clinical trials as well as new methods for stratified and personalised medicine. The estimates obtained by solving Equations (3), (4), and (7) give the overall, stratified, and personalised treatment effect respectively, i.e. the first ($\hat{\boldsymbol{\vartheta}}$) applies to all patients, the second ($\hat{\boldsymbol{\vartheta}}_g$) to a subgroup g, and the last ($\hat{\boldsymbol{\vartheta}}_k$) to a single patient k. Model-based recursive partitioning methods can be used for exploratory analysis of clinical trials. The intended use of these methods is to identify *whether treatment effect heterogeneity* exists in the given patient population and then to *generate hypotheses* based on potential findings. These hypotheses can then be used to *inform new clinical trials*.

Software for computing model-based trees, model-based forests and PALM trees is available in open source R packages. The `partykit` package provides all base functionalities. Package `model4you` provides a user-friendly interface for the application of model-based trees and forests for stratified and personalised medicine including visualisation functionality for easy interpretation and communication of models and results. PALM trees are available in the `palmtree` package. Development versions of all packages are available on the R development platform R-Forge[2]. The `partykit` package is also available on the *Comprehensive R Archive Network (CRAN)*[3].

3 Thesis summary

This thesis consists of four papers. Each paper deals with different aspects of model-based recursive partitioning for stratified or personalised medicine. The first paper introduces model-based trees as a method for stratified medicine; the second paper describes model-based forests and the computation of personalised models; the third paper introduces the PALM tree algorithm; The fourth paper describes the R package `model4you`, which implements model-based trees, model-based forests and personalised models in a user-friendly way. The contents are summarised below.

[2] https://r-forge.r-project.org/projects/partykit/
[3] https://CRAN.R-project.org/package=partykit

Paper I: Model-Based Recursive Partitioning for Subgroup Analyses

by Heidi Seibold, Achim Zeileis, and Torsten Hothorn

This article describes the basics of how model-based recursive partitioning can be used for stratified medicine. Model-based trees are used to analyse data collected in several clinical trials about Amyotrophic Lateral Sclerosis (ALS) and the effect of the drug Riluzole (the only approved drug against ALS) on survival and health of patients. The models used within the model-based trees are the following: Survival is analysed both by partitioning a Weibull model as well as a Cox model. Health is measured by a ten item sum score, the ALS Functional Rating Scale (ALSFRS), and analysed by partitioning a Gaussian model. Since not only the sum score but also each item is of interest, ten proportional odds models, one model per item, are combined in a tree by combining the score matrices of all ten models. The paper shows that model-based trees are able to identify predictive and prognostic factors and that the corresponding visualisations allow for easy interpretation, which makes them a good tool to communicate the results.

At the *Workshop on Classification and Regression Trees*[4] (March 2014), sponsored by the Institute for Mathematical Sciences of the National University of Singapore my supervisors, Torsten Hothorn and Achim Zeileis, learned about the need for methods for stratified medicine and realised that model-based trees are good for this use-case. This sparked the idea for my Ph.D. project and lead to this first paper. It was published in 2016 in the *International Journal of Biostatistics*.

Paper II: Individual treatment effect prediction for amyotrophic lateral sclerosis patients

by Heidi Seibold, Achim Zeileis, and Torsten Hothorn

This article explains how model-based forests can be used to estimate similarity of patients in terms of model parameters, e.g. intercept (or baseline hazard) and treatment effect, and, with the help of this similarity measure, estimate personalised models. Again the survival and health of ALS patients is analysed. Treatment of interest is again Riluzole (versus no treatment). The personalised models show that treatment heterogeneity is present and visualising personalised treatment effects shows interesting patterns for relevant patient characteristics. We investigated the performance of the method in simulations and showed that it works well even in scenarios with complicated treatment $\times$ covariate interactions.

The paper is published in the Journal *Statistical Methods in Medical Research*.

[4] http://www2.ims.nus.edu.sg/Programs/014swclass

Paper III: Generalised Linear Model Trees with Global Additive Effects

by Heidi Seibold, Torsten Hothorn, and Achim Zeileis

In this article we propose partially additive linear model (PALM) trees as a way to gain all benefits from model-based trees and additionally allowing for covariate effects that are the same for all subgroups. An extensive simulation study shows the performance of PALM trees in comparison to competitors including classical model-based trees. PALM trees perform well in scenarios with global prognostic factors and, at the same time, keep the ability of model-based trees to stick to the global model and do not split in case of no subgroups. We applied the method in a setting where performance of students in a mathematics exam is of interest and students receive one of two slightly different exams. Knowledge about their performance throughout the semester is known and an obvious prognostic factor that can be included in the model.

This project was inspired by Sies and Van Mechelen (2017). They compare different methods for the detection of optimal treatment regimes, including model-based trees, in a simulation study. The way they simulated data suggested that there are known prognostic factors that have a linear effect on the outcome of interest. Beyond the grant for my Ph.D. project (205321_163456) the Swiss National Science Foundation supported this project with a mobility grant (205321_163456/2) which allowed me to visit Achim Zeileis' group at the University of Innsbruck. The article is available on the pre-print server arXiv[5] and waiting for the second round of reviews at the Journal *Advances in Data Analysis and Classification*.

Paper IV: `model4you`: An R package for personalised treatment effect estimation

by Heidi Seibold, Achim Zeileis, and Torsten Hothorn

The final article introduces the R package `model4you`[6] which implements the methodology of papers I and II. The package focuses on ease of use and interpretability. It provides a very simple interface, where users compute the overall model (base model) and insert it in the `pmtree` or `pmforest` function to compute model-based trees and forests respectively. The `pmodel` function can be used to compute personalised models from the model-based forest results. It is important that the software is open source, easy to use and produces high quality visualisations, so that a broad audience can use our methods. The manuscript describes a simple use case and gives an overview of the software information.

[5] https://arxiv.org/abs/1612.07498

[6] The R package `model4you` is currently available on https://r-forge.r-project.org/projects/partykit/

I will submit the article to the Journal of Open Research Software as soon as the package is published on the *Comprehensive R Archive Network (CRAN)*.

4 Impact and outreach

Although only two of my four papers have been published, my work has already started to get attention and make an impact.

Model-based trees for stratified medicine as described in my first article were used by Sies and Van Mechelen (2017) in a simulation study, comparing tree-methods that are able to detect optimal treatment rules. The only competitive opponent was a method by Zhang et al. (2012), which can not estimate treatment effects as of itself, but only create treatment rules.

Beyond classical subgroup analysis, where the focus is on estimating treatment effects for a given treatment dose, subgroup analysis for dose-response models is one task in drug development. In a collaboration with Marius Thomas and Björn Bornkamp, statisticians at the pharmaceutical company Novartis, we implemented model-based trees for dose-response models and studied their behaviour in a simulation study and in a phase II clinical trial assessing the efficacy of a new treatment for an inflammatory disease. The results suggest that model-based trees can be used for dose-response subgroup analysis. The paper with the title "Subgroup identification in dose-finding trials via model-based recursive partitioning" was accepted for pulishing in the journal *Statistics in Medicine*.

Yi-Ping Lin, research associate at the Koo Foundation Sun Yat-Sen Cancer Hospital in Taiwan, uses model-based trees and forests to analyse breast cancer data collected at the hospital since 1990. The aim is to provide a website with recommendations for treatment of breast cancer patients based on patient characteristics.

For my second paper "Individual treatment effect prediction for amyotrophic lateral sclerosis patients", published in the journal *Statistical Methods in Medical Research* I received the *Arthur-Linder Prize*[7], which is awarded to a young member of the *Austro-Swiss Region of the International Biometric Society (RoeS)* every second year. The prize is awarded at the for an excellent research paper in the field of biometrics. I received the prize in 2017 at the *Joint Conference on Biometrics & Biopharmaceutical Statistics*[8] in Vienna.

In November 2016 I visited the German Cancer Research Centre following an invitation by Axel Benner. His group is working on high dimensional data and they

[7] https://www.ibs-roes.org/home-en/arthur-linder-prize
[8] http://cenisbs2017.org

are interested in using our methods in this context. We are currently collaborating on a project analysing the personalised treatment effects of patients suffering from acute myeloid leukemia (AML). Patient characterstics in this project are 19,656 gene expressions. The high dimensionality of the split variables demands a new infrastructure for the computation of personalised models and good computational performance. Julia Krzykalla is currently starting her Ph.D. project at the German Cancer Research Centre, extending on my research. Her tentative project title is: "Modelling strategies for the identification of prognostic and predictive factors in competing risks and multi-state models".

Simon Foster, psychologist at the University of Zurich, studies the treatment response of adolescents suffering from depression. Several treatments or treatment combinations are available. In a collaboration with Lynette Tay, Meichun Mohler-Kuo and Torsten Hothorn, we estimated personalised treatment responses for the *Treatment for Adolescents With Depression Study (TADS)* of all two-way treatment comparisons: cognitive-behavioural therapy versus Fluoxetine (an antidepressant drug, which inhibits selective serotonin reuptake); cognitive-behavioural therapy versus the combination of cognitive-behavioural therapy and Fluoxetine; and Fluoxetine versus the combination of cognitive-behavioural therapy and Fluoxetine. The results show that the combination of cognitive-behavioural therapy with Fluoxetine was consistently superior to either therapy alone across patients and should be the preferred treatment.

In November 2017 I was invited to present a the *Biogen Symposium on Statistical Methods in Multiple Sclerosis* at the Biogen Corporate Offices in Cambridge, Massachusetts. Biogen is interested in a collaboration to apply our methods for multiple sclerosis research.

Throughout my Ph.D. I had several opportunities to present this work at international conferences including the *useR! 2016* conference in Stanford (USA) and the *CEN ISBS* conference 2017 in Vienna (Austria). In 2017 I was an invited speaker at the *Statistical Computing* workshop in Günzburg (Germany). Slides for my presentations are available on GitLab[9].

References

Atassi, N., Berry, J., Shui, A., Zach, N., Sherman, A., Sinani, E., Walker, J., Katsovskiy, I., Schoenfeld, D., Cudkowicz, M. and Leitner, M. (2014). The PRO-ACT database: Design, initial analyses, and predictive features, *Neurology* **83**(19): 1719–1725. doi:10.1212/WNL.0000000000000951.

Breiman, L. (2001). Random forests, *Machine Learning* **45**(1): 5–32. doi:10.1023/A:1010933404324.

[9] https://gitlab.com/research_heidi_seibold/slides/wikis/home

Das, R. (2017). Drug industry bets big on precision medicine: Five trends shaping care delivery. URL: https://www.forbes.com/sites/reenitadas/2017/03/08/drug-development-industry-bets-big-on-precision-medicine-5-top-trends-shaping-future-care-delivery.

Dusseldorp, E. and Van Mechelen, I. (2013). Qualitative interaction trees: A tool to identify qualitative treatment-subgroup interactions, *Statistics in Medicine* **33**(2): 219–237. doi:10.1002/sim.5933.

European Medicines Agency (2012). Riluzole Zentiva: EPAR summary for the public. URL: http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Summary_for_the_public/human/002622/WC500127609.pdf.

European Medicines Agency (2014). EMA guideline on the investigation of subgroups in confirmatory clinical trials (draft). URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/02/WC500160523.pdf.

Foster, J. C., Taylor, J. M. and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data, *Statistics in Medicine* **30**(24): 2867–2880. doi:10.1002/sim.4322.

Frueh, F. W. and Gurwitz, D. (2004). From pharmacogenetics to personalized medicine: a vital need for educating health professionals and the community, *Pharmacogenomics* **5**(5): 571–579. doi:10.1517/14622416.5.5.571.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*, Springer Series in Statistics, 2nd edn, Springer.

Italiano, A. (2011). Prognostic or predictive? It's time to get back to definitions!, *Journal of Clinical Oncology* **29**(35): 4718–4718. doi:10.1200/JCO.2011.38.3729.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction, *Econometric Theory* **21**(01): 21–59. doi:10.1017/S0266466605050036.

Negassa, A., Ciampi, A., Abrahamowicz, M., Shapiro, S. and Boivin, J. F. (2005). Tree-structured subgroup analysis for censored survival data: Validation of computationally inexpensive model selection criteria, *Statistics and Computing* **15**(3): 231–239. doi:10.1007/s11222-005-1311-z.

New York Times Editorial Board (2015). Medicine just for you. URL: https://www.nytimes.com/2015/02/08/opinion/sunday/medicine-just-for-you.html.

Schleidgen, S., Klingler, C., Bertram, T., Rogowski, W. H. and Marckmann, G. (2013). What is personalized medicine: sharpening a vague term based on a systematic literature review, *BMC Medical Ethics* **14**(1): 55. doi:10.1186/1472-6939-14-55.

Seibold, H., Hothorn, T. and Zeileis, A. (2016). Generalised linear model trees with global additive effects, *ArXiv e-prints* . URL: https://arxiv.org/abs/1612.07498.

Seibold, H., Zeileis, A. and Hothorn, T. (2016). Model-based recursive partitioning for subgroup analyses, *International Journal of Biostatistics* **12**(1): 45–63. doi:10.1515/ijb-2015-0032.

Seibold, H., Zeileis, A. and Hothorn, T. (2017a). Individual treatment effect prediction for amyotrophic lateral sclerosis patients, *Statistical Methods in Medical Research* . online first. doi:10.1177/0962280217693034.

Seibold, H., Zeileis, A. and Hothorn, T. (2017b). model4you: An R package for personalised treatment effect estimation.

Sies, A. and Van Mechelen, I. (2017). Comparing four methods for estimating tree-based treatment regimes, *The International Journal of Biostatistics* **Online First**. doi:10.1515/ijb-2016-0068.

Slamon, D. J., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., Bajamonde, A., Fleming, T., Eiermann, W., Wolter, J., Pegram, M., Baselga, J. and Norton, L. (2001). Use of chemotherapy plus a monoclonal antibody against her2 for metastatic breast cancer that overexpresses her2, *New England Journal of Medicine* **344**(11): 783–792. doi:10.1056/NEJM200103153441101.

US Food and Drug Administration (2013). Paving the way for personalized medicine. URL: http://www.fda.gov/downloads/ScienceResearch/SpecialTopics/PersonalizedMedicine/UCM372421.pdf.

Weisberg, H. I. (2015). What next for randomised clinical trials?, *Significance* **12**(1): 22–27. doi:10.1111/j.1740-9713.2015.00798.x.

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M. and Laber, E. (2012). Estimating optimal treatment regimes from a classification perspective, *Stat* **1**(1): 103–114. doi:10.1002/sta.411.

PAPER I

Model-Based Recursive Partitioning for Subgroup Analyses

Heidi Seibold, Achim Zeileis, Torsten Hothorn

Published in *The International Journal of Biostatistics*, 2016, **12** (1), 45–63.

Open Access

Heidi Seibold, Achim Zeileis and Torsten Hothorn*

Model-Based Recursive Partitioning for Subgroup Analyses

DOI 10.1515/ijb-2015-0032

Abstract: The identification of patient subgroups with differential treatment effects is the first step towards individualised treatments. A current draft guideline by the EMA discusses potentials and problems in subgroup analyses and formulated challenges to the development of appropriate statistical procedures for the data-driven identification of patient subgroups. We introduce model-based recursive partitioning as a procedure for the automated detection of patient subgroups that are identifiable by predictive factors. The method starts with a model for the overall treatment effect as defined for the primary analysis in the study protocol and uses measures for detecting parameter instabilities in this treatment effect. The procedure produces a segmented model with differential treatment parameters corresponding to each patient subgroup. The subgroups are linked to predictive factors by means of a decision tree. The method is applied to the search for subgroups of patients suffering from amyotrophic lateral sclerosis that differ with respect to their Riluzole treatment effect, the only currently approved drug for this disease.

Keywords: subgroup analysis, personalized medicine, treatment efficacy, permutation test, amyotrophic lateral sclerosis

1 Introduction

With the rise of personalised medicine, the search for individual treatments poses challenges to the development of appropriate statistical methods. Subgroup analyses following a traditional statistical assessment of an overall treatment effect of a new therapy aim at identifying three groups of patients: (1) those who benefit from the new therapy, (2) those who do not benefit, and (3) those whose clinical outcome under the new therapy is worse than under alternative therapies. Such post-hoc subgroup analyses potentially lead to better benefit-risk decisions and treatment recommendations but are subject to all kind of biases and can hardly be performed under full statistical error control. Therefore, the European Medicines Agency (EMA) recently published a draft of a guideline for the investigation of subgroups in confirmatory clinical trials [1] that discusses potential areas of application, necessity, pitfalls, and good practice in subgroup analyses. In the guideline draft, three scenarios in which exploratory investigation of subgroups is of special interest were identified:

Scenario 1: "The clinical data presented are overall statistically persuasive with therapeutic efficacy demonstrated globally. It is of interest to verify that the conclusions of therapeutic efficacy (and safety) apply consistently across subgroups of the clinical trial population."

Scenario 2: "The clinical data presented are overall statistically persuasive but with therapeutic efficacy or benefit/risk which is borderline or unconvincing and it is of interest to identify post-hoc a subgroup, where efficacy and risk-benefit is convincing."

Scenario 3: "The clinical data presented fail to establish statistically persuasive evidence but there is interest in identifying a subgroup, where a relevant treatment effect and compelling evidence of a favourable risk-benefit profile can be assessed."

***Corresponding author: Torsten Hothorn,** Department of Biostatistics, Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland

Heidi Seibold, Department of Biostatistics, Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland

Achim Zeileis, Department of Statistics, Faculty of Economics and Statistics University of Innsbruck, Innsbruck, Austria

Especially in trials with highly heterogeneous study populations, subgroup analyses can help to reduce the variability of the estimated overall treatment effect by splitting the study population into more homogeneous subgroups.

Information about the individual treatment effect might be available from cross-over trials or from counterfactual analyses of parallel-group designs [2, 3]. These individual effects can then be linked to potentially predictive variables. In the absence of such information, most importantly in the case of parallel-group designs studied here, subgroup analyses can be seen as the search for or specification of treatment × covariate interactions and we proceed along this path. A covariate measures a patient characteristic that potentially explains the patient's individual treatment effect. In the commonly applied models with linear predictors, such as the linear, generalised linear or linear transformation models, the specification of higher-order interaction terms and especially the subsequent inference are known to be burdensome. For non-categorical covariates, it is a priori unclear how one can derive a subgroup from a significant treatment × covariate interaction.

Automated interaction detection [4], today known as recursive partitioning methods or simply "trees", was suggested as an interaction search procedure more than 50 years ago, and has had a very active development community ever since. Although the application of trees for subgroup identification seems to be straightforward, no generally applicable method is available [5]. The main technical problem is that classical trees were developed for identifying higher-order covariate interactions but additional work is required to restrict interactions to treatment × covariate interactions. Due to the non-parametric nature of most tree models, blending trees with the linear models typically used to describe the treatment effect is challenging.

While setting up such automated procedures for subgroup identification, one has to bear in mind that the impact of a covariate on the endpoint can be prognostic, predictive, or both. Prognostic factors have a direct impact on the endpoint, independent of the treatment applied. This corresponds to a main effect. A predictive factor explains a differential treatment effect, i.e. a treatment × covariate interaction term. Both the main and the treatment interaction terms are important for factors that are prognostic and predictive at the same time [6].

In our analysis, we aimed at detecting subgroups of patients suffering from amyotrophic lateral sclerosis (ALS) in which the subgroups differ in the effect of treatment with Riluzole, the only approved drug for ALS treatment today. The two endpoints of interest are a functional endpoint assessing the patient's ability to handle daily life and the overall survival time. We estimated the overall treatment effect of Riluzole using four different base models; the choice of the model depended on the measurement scale of the endpoint. A normal generalised linear model (GLM) with log-link was used for the sum-score of the functional endpoint, and item-specific proportional odds models were used for the decomposed score. For the right-censored survival times, we used a parametric Weibull model and a semiparametric Cox model. Our aim was to partition these linear models with respect to the treatment effect parameter and to develop a segmented model that includes treatment × covariate interactions that describe the relevant subgroups.

We applied model-based recursive partitioning [7] to the functional and survival models describing the effect of Riluzole on ALS patients in order to obtain subgroups with a differential treatment effect. The main advantage of embedding our subgroup analysis into this general framework of model partitioning is that one can partition the base model used for analysing the overall treatment effect, regardless of the measurement scale of the endpoint. The method allows us to focus attention on predictive factors, while other terms, such as the effects of strata or nuisance parameters, can be held fixed.

Section 2 introduces the general framework for subgroup identification and compares the new procedure to methods published previously in the light of this general theoretical framework. In Section 3, we present results of our subgroup analysis of Riluzole treatment of ALS patients and discuss the patient subgroups and corresponding differential treatment effects found.

2 Model-based recursive partitioning for subgroup identification

Subgroup analyses require the definition of a parameter describing the treatment effect. In clinical trials, this parameter is typically already contained in the model that was defined in the study protocol for the

analysis of the primary endpoint. The treatment effect was estimated in the primary analysis under the assumption that the corresponding parameter is universally applicable to all patients. In the presence of subgroups, this assumption does not hold and these patient subgroups differ in their treatment effect. If we assume that the different treatment effects can be understood as a function of patient characteristics, the patient subgroups can be identified by estimating this treatment effect function. Model-based recursive partitioning can be employed as a procedure for the estimation of such a treatment effect function and the identification of the corresponding patient subgroups. The name of the procedure comes from the nature of the algorithm that recursively partitions the initial model used for the analysis of the primary endpoint.

2.1 Model and algorithm

We started with a model $\mathscr{M}((Y, \mathbf{X}), \vartheta)$ that describes the conditional distribution of the primary endpoint Y (or certain characteristics of this distribution) as a function of the treatment arm and potentially further covariates (both contained in $\mathbf{X}$) through parameters ϑ as defined in the study protocol. The parameter vector $\vartheta = (\alpha, \beta, \gamma, \sigma)^\top$ typically contains one or more intercept parameters α, one or more treatment effect parameters β, other model parameters γ, e.g. effects of covariates, and potential nuisance parameters σ, e.g. the error variance in a linear model. The estimator is defined as the minimizer of an objective function Ψ, which usually is the negative log-likelihood:

$$\hat{\vartheta} = \arg\min_{\vartheta} \sum_{i=1}^{N} \Psi((y, x)_i, \vartheta). \tag{1}$$

Estimating ϑ is equivalent to solving the score equation

$$\sum_{i=1}^{N} \frac{\partial \Psi((y, \mathbf{x})_i, \vartheta)}{\partial \vartheta} = \sum_{i=1}^{N} \psi((y, \mathbf{x})_i, \vartheta) = 0, \tag{2}$$

where ψ is the score function, i.e. the gradient of the objective function Ψ with respect to ϑ. The model framework is more general than the log-likelihood framework because Ψ is not necessarily a negative log-likelihood function.

In the presence of patient subgroups that differ in their treatment effect β, an estimate $\hat{\beta}$ obtained for all patients $i = 1, \ldots, N$ in the study only reflects the mean treatment effect but ignores that the success or failure of a specific treatment might depend on additional characteristics of each individual patient. We describe patient subgroups as a partition $\{\mathscr{B}_b\}$ $(b = 1, \ldots, B)$ of all patients $i = 1, \ldots, N$. The subgroup-specific model parameters are then $\vartheta(b)$. These parameters can in general be seen as varying coefficients [8], however they may depend on several patient characteristics and are always step functions with a different level for each subgroup and not only a smoothly varying coefficient for one single predictive variable.

Since we are searching for predictive and prognostic factors, we are only interested in subgroups that differ in the intercept or the treatment effect or both as explained in Section 2.2. With $\vartheta(b) = (\alpha(b), \beta(b), \boldsymbol{\gamma}, \sigma)^\top$ we assume that the effects of covariates and nuisance parameters are constant for all patients. The partition $\{\mathscr{B}_b\}$ is defined by J partitioning variables $\mathbf{Z} = (Z_1, \ldots, Z_J) \in \mathscr{Z}$; in other words, $\{\mathscr{B}_b\}$ is a hypercube in the J-dimensional sample space $\mathscr{Z}$. These partitioning variables $\mathbf{Z}$ are the additional patient characteristics that potentially influence $\alpha(b)$ and $\beta(b)$. If for example gender were a predictive factor in a given treatment-endpoint relationship, it would be a patient characteristic that is involved in forming the partitions. If the partition $\{\mathscr{B}_b\}$ is known, the partitioned model parameters $\vartheta(b)$ could be estimated by minimising the segmented objective function:

$$(\hat{\vartheta}(b))_{b=1,\ldots,B} = \arg\min_{\vartheta(b)} \sum_{i=1}^{N} \sum_{b=1}^{B} \mathbb{1}(\mathbf{z}_i \in \mathscr{B}_b) \Psi((y, \mathbf{x})_i, \vartheta(b)), \tag{3}$$

where $\mathbb{1}$ denotes the indicator function and $(y, \mathbf{x})_i, \mathbf{z}_i$ are the realisations of $(Y, \mathbf{X})$ and $\mathbf{Z}$ for the i-th patient. This allows us to write the subgroup-specific intercept and treatment parameters as a function of the partitioning variables

$$\alpha(\mathbf{z}) = \sum_{b=1}^{B} \mathbb{1}(\mathbf{z} \in \mathscr{B}_b) \cdot \alpha(b) \quad \text{and} \quad \beta(\mathbf{z}) = \sum_{b=1}^{B} \mathbb{1}(\mathbf{z} \in \mathscr{B}_b) \cdot \beta(b).$$

Without any a priori knowledge about the partition $\{\mathscr{B}_b\}$, we want to estimate the functions $\alpha(\mathbf{z})$ and $\beta(\mathbf{z})$ by means of model-based recursive partitioning. The main idea underlying this method is the ability to detect parameter instabilities, i.e. non-constant parameters in a parametric or semiparametric model, by looking at the score function. Because we are only interested in detecting non-constant intercepts $\alpha(\mathbf{z})$ and treatment effects $\beta(\mathbf{z})$, we focus on the partial score functions $\psi_\alpha((Y, \mathbf{X}), \vartheta) = \partial\Psi((Y, \mathbf{X}), \vartheta)/\partial\alpha$ and $\psi_\beta((Y, \mathbf{X}), \vartheta) = \partial\Psi((Y, \mathbf{X}), \vartheta)/\partial\beta$. If the model parameters are in fact constant and do not depend on any of the partitioning variables $\mathbf{Z}$, the partial score functions $\psi_\alpha((Y, \mathbf{X}), \vartheta)$ and $\psi_\beta((Y, \mathbf{X}), \vartheta)$ are independent of $\mathbf{Z}$. Consequently, parameter instability corresponds to a correlation between either of the partial score functions and at least one of the partitioning variables $Z_1, \ldots, Z_J$. In order to formally detect deviations from independence between the partial score functions and the partitioning variables, model-based recursive partitioning utilises independence tests. The null hypotheses

$$H_0^{\alpha,j}: \quad \psi_\alpha((Y, \mathbf{X}), \hat{\vartheta}) \quad \perp \quad Z_j, j = 1, \ldots, J$$
$$\text{and}$$
$$H_0^{\beta,j}: \quad \psi_\beta((Y, \mathbf{X}), \hat{\vartheta}) \quad \perp \quad Z_j, j = 1, \ldots, J$$

for a given model $\mathscr{M}((Y, \mathbf{X}), \hat{\vartheta})$ state that the partial score functions with respect to α and β, respectively, are independent of the partitioning variable Z_j $(j = 1, \ldots, J)$. Hence, these null hypotheses correspond to an appropriate model fit regarding the intercept and treatment parameter. Because the partial score functions under the null hypotheses are at least asymptotically normal in many model families, asymptotic M-fluctuation tests with appropriate correction for multiplicity were introduced for model-based recursive partitioning by Zeileis and coworkers [9, 7]. Alternatively, permutation tests can be applied in situations where asymptotic normality of the partial score is not guaranteed [10] or in cases with small numbers of observations [11–13], which are common in medicine. Also in this case procedures for multiple testing are used to cope with a possibly large number of partitioning variables J.

If we can reject at least one of the $2 \times J$ null hypotheses for the global model $\mathscr{M}((Y, \mathbf{X}), \hat{\vartheta})$ at a pre-specified nominal level, model-based recursive partitioning selects the partitioning variable Z_{j^*} associated with the highest correlation to any of the partial score functions. This is usually done by means of the smallest p-value. The dependency structure between the partitioning variable Z_{j^*} and either one of the partial score functions is described by a simple cut-point model. Once we find an optimal cut-point $Z_{j^*} < \mu$ using a suitable criterion [13, 7], we split the patients into two subgroups according to $Z_{j^*} < \mu$. For both subgroups, we estimate two separate models with parameters $\hat{\vartheta}(1)$ and $\hat{\vartheta}(2)$, respectively, obtain the corresponding partial score functions, and test the independence hypotheses. If we find deviations from independence, we in turn estimate a cut-point in the most highly associated partitioning variable, and split again. The procedure of testing independence of partial score functions and partitioning variables is repeated recursively until deviations from independence can no longer be detected.

Since model-based recursive partitioning is a tree method, in the following we use topic-specific vocabulary, such as nodes. The root node contains all patients and is the basis for the initial model, inner nodes represent splits and leaf nodes contain the patients of the different subgroups and specify the partition-specific models. The paths from root to leaf nodes define the subgroups.

2.2 Content interpretation

A clearer picture of the interpretation of subgroup-dependent model parameters and distribution of the partial scores under unstable parameters is best given by means of a partitioned linear model discussed in the following.

Here x_A is a contrast that indicates whether a subject was treated with treatment A (active) but not C (control) in a two-armed trial and x_{stratum} is a stratum with $\mathbf{x} = (x_A, x_{\text{stratum}})$. The conditional distribution of the primary endpoints Y given treatment and stratum is normal

$$Y|\mathbf{X} = \mathbf{x} \sim \mathcal{N}(\alpha + \beta x_A + \gamma x_{\text{stratum}}, \sigma^2). \tag{4}$$

The segmented model we want to fit using model-based recursive partitioning reads

$$Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z} \sim \mathcal{N}(\alpha(\mathbf{z}) + \beta(\mathbf{z}) x_A + \gamma x_{\text{stratum}}, \sigma^2), \tag{5}$$

where γ is the effect of the stratum and the variance σ^2 is a nuisance parameter. The objective function for a patient with observations $(y, \mathbf{x})$ is the negative log-likelihood, when maximum likelihood estimation is used, or the error sum of squares, when ordinary least squares is used. Yet, both methods lead to the same scores

$$\psi((y, \mathbf{x}), \hat{\vartheta}) = \begin{pmatrix} \frac{\partial \Psi((y,\mathbf{x}),\boldsymbol{\vartheta})}{\partial \alpha}\Big|_{\boldsymbol{\vartheta}=\hat{\vartheta}} \\ \frac{\partial \Psi((y,\mathbf{x}),\boldsymbol{\vartheta})}{\partial \beta}\Big|_{\boldsymbol{\vartheta}=\hat{\vartheta}} \end{pmatrix}^\top = \frac{1}{\sigma^2} \begin{pmatrix} y - (\hat{\alpha} + \hat{\beta} x_A + \hat{\gamma} x_{\text{stratum}}) \\ (y - (\hat{\alpha} + \hat{\beta} x_A + \hat{\gamma} x_{\text{stratum}})) \cdot x_A \end{pmatrix}^\top \tag{6}$$

and thus to the same solution. Note that the partial score function with respect to the intercept is proportional to the least-square residuals and all further scores are proportional to the product of the residuals and the respective variable.

A partitioning variable can be predictive, prognostic, or both, and we have to consider the parameters in the model to understand the nature of a partitioning variable. Figure 1 shows examples for mean primary endpoints and the corresponding intercept α and treatment effect β. If $\alpha(\mathbf{z})$ varies over $\mathbf{z}$, but $\beta(\mathbf{z})$ is constant, then the components of $\mathbf{z}$ are prognostic because the mean primary endpoint varies but not the treatment effect (see first column of Figure 1). If $\beta(\mathbf{z})$ varies over $\mathbf{z}$ and $\alpha(\mathbf{z})$ is constant, then the variables in $\mathbf{z}$ are predictive since it means that the mean primary endpoint in one treatment arm stays the same but the treatment effect changes over $\mathbf{z}$ (second column). If both parameters vary, then $\mathbf{z}$ is predictive (third column) or predictive and prognostic at the same time (last column). In the latter situation, the mean primary endpoint of the second subgroup changes over $\mathbf{z}$ and the intercept also changes.

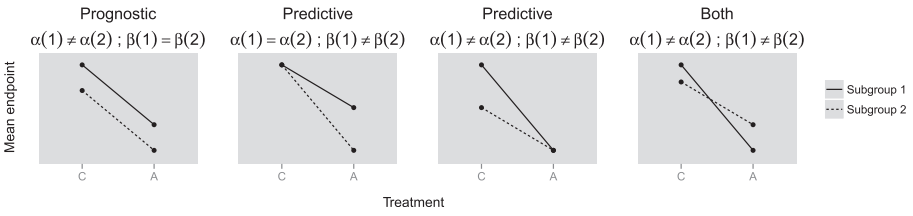


Figure 1: Possible mean primary endpoint within subgroups resulting from a predictive, prognostic, or predictive and prognostic variable.

It is also interesting to take a closer look at the partial scores. Figure 2a shows the partial scores with respect to intercept and treatment parameter that result from a linear model $Y|\mathbf{X} = \mathbf{x} \sim \mathcal{N}(\alpha + \beta x_A,\ \sigma^2)$ plotted against a partitioning variable z_1, which is predictive and prognostic. The data-generating process of this model was suggested by Loh et al. [14] and is defined as

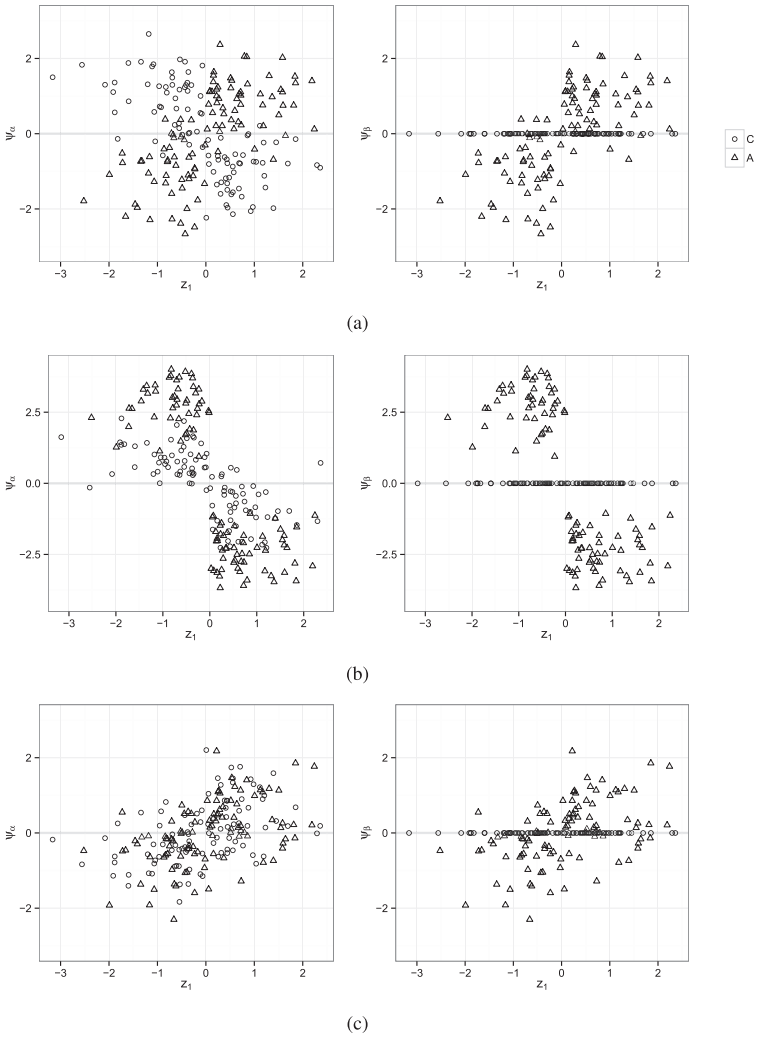


Figure 2: Partial scores of different kinds of variables. The symbols represent the treatment arms C and A as indicated. (a) Partial scores of a predictive and prognostic variable (eq. (7)). (b) Partial scores of a predictive and prognostic variable (eq. (8)). (c) Partial scores of a prognostic variable (eq. (9)).

$$Y|\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z} \sim \mathcal{N}(1.9+0.2\cdot x_A+1.8\cdot \mathbb{1}(z_1<0)+3.6\cdot \mathbb{1}(z_1>0)\cdot x_A, 0.7), \tag{7}$$

with X_A from $\mathcal{B}(1, 0.5)$ and Z_1 from $\mathcal{N}(0,1)$. For the example, we used this process to draw a sample of 200 observations.

The partial scores with respect to the intercept ψ_α fluctuate randomly around zero over the whole range of z_1. The partial scores with respect to the treatment parameter ψ_β change. Hence, in this situation, model-based recursive partitioning would detect a deviation from independence between ψ_β and z_1 and implement a split at approximately $z_1 < 0$. There is no chance of finding this cut-point by looking at the least-square residuals only, since a deviation of independence between ψ_α and z_1 is hardly visible in the scatterplot in

the left panel of Figure 2a. Figure 2b shows the partial scores obtained with a slightly modified data-generating process, where instead of $\mathbb{1}(z_1 > 0) \cdot x_A$, one has $\mathbb{1}(z_1 < 0) \cdot x_A$:

$$Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z} \sim \mathcal{N}(1.9 + 0.2 \cdot x_A + 1.8 \cdot \mathbb{1}(z_1 < 0) + 3.6 \cdot \mathbb{1}(z_1 < 0) \cdot x_A, 0.7). \tag{8}$$

Here the procedure would split the partial score with respect to the intercept, although z_1 is still prognostic and predictive at the same time.

If we focus on the prognostic variable z_1 in the model

$$Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z} \sim \mathcal{N}(2 \cdot x_A + \mathbb{1}(z_1 > 0), 0.7), \tag{9}$$

we see non-random patterns in both scores (see Figure 2c). Since the partial scores with respect to the treatment parameter are set to zero for treatment arm A, we would split on basis of the scores with respect to the intercept, just as a consequence of a higher power.

These three examples show that splitting in the partial score with respect to the intercept does not give any information about whether the partitioning variable is predictive or prognostic. It also does not make sense to choose to split only in the score with respect to the treatment parameter because one might miss important cut-points. In order to be able to say whether a partitioning variable is predictive or prognostic, it is not enough to know which partial scores are responsible for the split. It is necessary to consider the model parameters in the segmented model. If the treatment parameter β varies in the subgroups, then the chosen partitioning variables are predictive or both predictive and prognostic. If β is constant, the variables are only prognostic.

2.3 Relation to established procedures

Traditional approaches for subgroup identification are also based on a model for the primary endpoint, but the segmentation is implemented by means of varying coefficients. More precisely, the model includes interactions between treatment and the patient characteristics $\mathbf{z}$ in addition to the main effects

$$\begin{aligned} \mathbb{E}(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) &= \alpha + \beta x_A + \gamma_{\text{prognostic}}^{\mathrm{T}} \mathbf{z} + \gamma_{\text{predictive}}^{\mathrm{T}} \mathbf{z} x_A \\ &= (\alpha_0 + \gamma_{0,z}^{\mathrm{T}} \mathbf{z})(1 - x_A) + (\alpha_1 + \gamma_{1,z}^{\mathrm{T}} \mathbf{z}) x_A, \end{aligned} \tag{10}$$

with $\alpha = \alpha_0, \beta = \alpha_1 - \alpha_0, \gamma_{\text{prognostic}}^{\mathrm{T}} = \gamma_{0,z}^{\mathrm{T}}$ and $\gamma_{\text{predictive}}^{\mathrm{T}} = \gamma_{1,z}^{\mathrm{T}} - \gamma_{0,z}^{\mathrm{T}}$. The model is known as the "classical approach" for subgroup analyses [15, 16]. Significant interaction terms $\gamma_{\text{predictive}}$ are in this case subject to the choice of relevant partitioning variables. However, patient subgroups can only be identified directly in this model for categorical variables z_j since the model has no notion of optimal cut-off points. As the number of potential partitioning variables J might be large, the simultaneous estimation of all parameters in the model might be computationally burdensome and associated with a large variance. Regularisation procedures may be applied for selecting relevant interaction parameters that deviate considerably from zero.

RECPAM [17, 18] goes a step further and fits such models by trees. In every node, a likelihood-ratio test is computed that compares the segmented model

$$\mathbb{E}(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = \alpha + \beta_1 x_A \mathbb{1}(z_j \in \mathscr{B}_k) + \beta_2 x_A [1 - \mathbb{1}(z_j \in \mathscr{B}_k)] \tag{11}$$

to the constant model

$$(\mathbb{E}Y|\mathbf{X} = \mathbf{x}) = \alpha + \beta x_A \tag{12}$$

for every possible segment $\mathscr{B}_k$ ($k = 1, \ldots, K$) induced by all possible cut-off points in z_j, i.e. an exhaustive search is performed. The procedure is applied to all partitioning variables z_j ($j = 1, \ldots, J$). The algorithm then chooses the variable and segmentation that comes along with the highest test statistic. The method is so far limited to linear models and Cox proportional hazards models, and parameter instabilities can only be detected in β but not in α.

A method that is similar in spirit to model-based recursive partitioning, but is limited to normal linear models, is the Gs method [14] based on the GUIDE algorithm [19, 20]. Instead of using partial scores with respect to intercept and treatment effect, Gs uses only the least-square residuals (that is, only the partial score with respect to the intercept). In contrast to model-based recursive partitioning, Gs looks at the dichotomised (at zero) residuals separately in the two treatment arms. The independency between positive/negative residual signs and each partitioning variable is tested using a chi-squared test separately for each treatment. If the partitioning variable is at least ordinal, it is dichotomised by splitting at the mean. The optimal split variable chosen is the one that induces the highest sum of chi-squared statistics. Looking at the left panels of Figures 2a and 2b, one can imagine that in these situations the procedure may successfully find the subgroups. However, in a less clear situation and where the optimal cut-point is not near the mean of z_1, the method will have lower power or will not be able to find a split at all.

Another recently proposed tree algorithm is qualitative interaction trees (QUINT [21]). QUINT searches for instabilities in the treatment parameter β only, but the resulting partitions have to have different signs in the parameter. In other words, QUINT aims at finding subgroups in which the treatment effect is the reverse of that of the other subgroups. The current implementation of QUINT [22] is limited to continuous primary endpoints. It would be possible to enforce splits that are qualitatively different in model-based recursive partitioning. This could be achieved by incorporating a criterion that implements a split only if the treatment effects in the two new subgroups have different signs.

SIDES (subgroup identification based on differential effect search) [23] and SIDEScreen [24] aim at identifying subgroups of patients with high benefit from a novum compared to the standard treatment. Although the subgroups are linked to hypercubes in the sample space of Z, they are overlapping and can therefore not be represented as a tree structure. The methods are based on a cross-validated implementation of subgroups that were obtained on independent learning samples.

More general approaches blending recursive partitioning with traditional models (known as hybrid, model, or functional trees in machine learning Gama [25]), include M5 [26], GUIDE [19], CRUISE [27], LOTUS [28] and maximum likelihood trees [29]. (Bayesian approaches can be found in Chipman, George, and McCulloch [30]) and (Bernardo, Bayarri, Berger, Dawid, Heckerman, Smith, and West [31]). Except GUIDE, none of these methods has been studied in the specific context of subgroup analyses so far.

3 Partitioning effects of Riluzole on ALS patients

ALS is a neurodegenerative disease that causes weakness, muscle waste and paralysis. Currently the only drug on the market for treating ALS is Riluzole (Rilutek). It slows down disease progression but only modestly prolongs life expectancy by about two months [32]. A more thorough investigation of the treatment effect of Riluzole in ALS patients is of great importance since a cure is not yet available and patients usually die within 1.5–4 years after disease onset [33]. We use model-based recursive partitioning to address the question whether Riluzole has an especially low or high treatment effect on both functional and survival endpoints of any subgroups of patients.

Our analysis is based on patient information obtained from the PRO-ACT (Pooled Resource Open-Access ALS Clinical Trials) database [34], which contains data of ALS patients that were involved in one of several publicly- and privately-conducted clinical trials. The database provides information on patient survival, functional endpoint (the ALS functional rating scale), Riluzole use, demographics, family history, patient history, forced and slow vital capacity, laboratory data and vital signs. The data were fully de-identified and therefore the centres of data ascertainment are not given in the data set. The participants gave their informed consent, and study protocols were approved in the respective medical centres. The database was initiated by the non-profit organisation Prize4Life that aims at accelerating cure and drug development for ALS, for example through the DREAM-Phil Bowen ALS Prediction Prize4Life challenge [35].

The ALS Functional Rating Scale (ALSFRS [36]), is a widely used instrument for evaluating the functional status of patients with ALS even though the uni-dimensionality of the score seems questionable [37]. It is a sum-score of the following ten items: speech, salivation, swallowing, handwriting, cutting food and handling utensils, dressing and hygiene, turning in bed and adjusting bed clothes, walking, climbing stairs, and breathing. Each of these items can have values from zero to four, where four is normal and zero indicates the inability of performing the respective action. Hence, if the ALSFRS has a value 40, the patient has normal abilities for all items. The lower the score, the worse is the patient's status. The items were measured at several time points during the study period. We focused on the ALSFRS reading approximately six months after treatment start as the functional endpoint. Approximately means that we used the measurement closest to six months after treatment start, with a maximal absolute deviation of 20 days. In addition, we also decomposed the score and modelled the items defining the score separately.

The survival time of patients was measured in days starting with the patient's enrolment in one of the trials. For patients without survival information, we used the latest follow-up time given for the patient in the data as censoring time.

Model-based recursive partitioning was applied to models for the functional and survival endpoints. We allowed parameter instabilities in both the intercept and the Riluzole treatment effect. Bonferroni-adjusted permutation tests using test statistics of a quadratic form [13] were applied for assessing independence of the partial score functions and the partitioning variables and also for cut-point selection. The use of permutation tests for cut-point selection improves speed compared to the original suggestion of fitting and comparing models for all reasonable partitions [7]. We restricted the depth of the trees to two levels. Parameter estimates including confidence intervals are given for the final subgroups. Note that we are computing the confidence intervals after applying model selection through splitting into subgroups and thus the intervals should be interpreted with caution. For both endpoints, we used partitioning variables available at patient enrolment from the following groups of variables: demographics, family history, patient history, forced and slow vital capacity, laboratory data, and vital signs. We excluded patient records with missing values at the endpoints; the sample size was $N = 2534$ for the functional endpoint and $N = 3306$ with 916 events for the survival endpoint.

3.1 ALSFRS

The ALSFRS six months after treatment start (ALSFRS_6) defined the functional endpoint. The sum-score is positive, and the model needs to adjust for the baseline ALSFRS obtained at treatment start (ALSFRS_0). We used a normal GLM with log-link and offset $log(\text{ALSFRS}_0)$ such that the model

$$\mathbb{E}\left(\frac{\text{ALSFRS}_6}{\text{ALSFRS}_0}|X = x\right) = \frac{\mathbb{E}(\text{ALSFRS}_6|X = x)}{\text{ALSFRS}_0} = \exp\{\alpha + \beta x_R\} \tag{13}$$

describes the expected relative change in the ALSFRS over the first six months under treatment. The treatment (Riluzole/no Riluzole) is indicated by x_R. The model was fitted by maximum likelihood.

The time between disease onset and start of treatment, the forced vital capacity (FVC), and the phosphorus balance are the three partitioning variables selected for the tree given in Figure 3. The FVC value gives the volume of air in liters that can forcibly be blown out after full inspiration to the lung. A normal phosphorus balance is between 1 and 1.5 mmol/L. The tree indicates a negative treatment effect of Riluzole for patients with fewer days between disease onset and start of treatment that have a higher FVC value (node 4). Therefore, the FVC value is predictive in the group of patients with less than 468 days between disease onset and treatment start. Patients with more days between disease onset and treatment start do not seem to have a treatment effect. The fact that the time since onset plays an important role is not surprising since it is a surrogate for the speed of disease progression [38]. Patients with a slow progression were seldom included early in one of the studies. Hence a long time between onset and start of treatment usually stands for a slow progression.

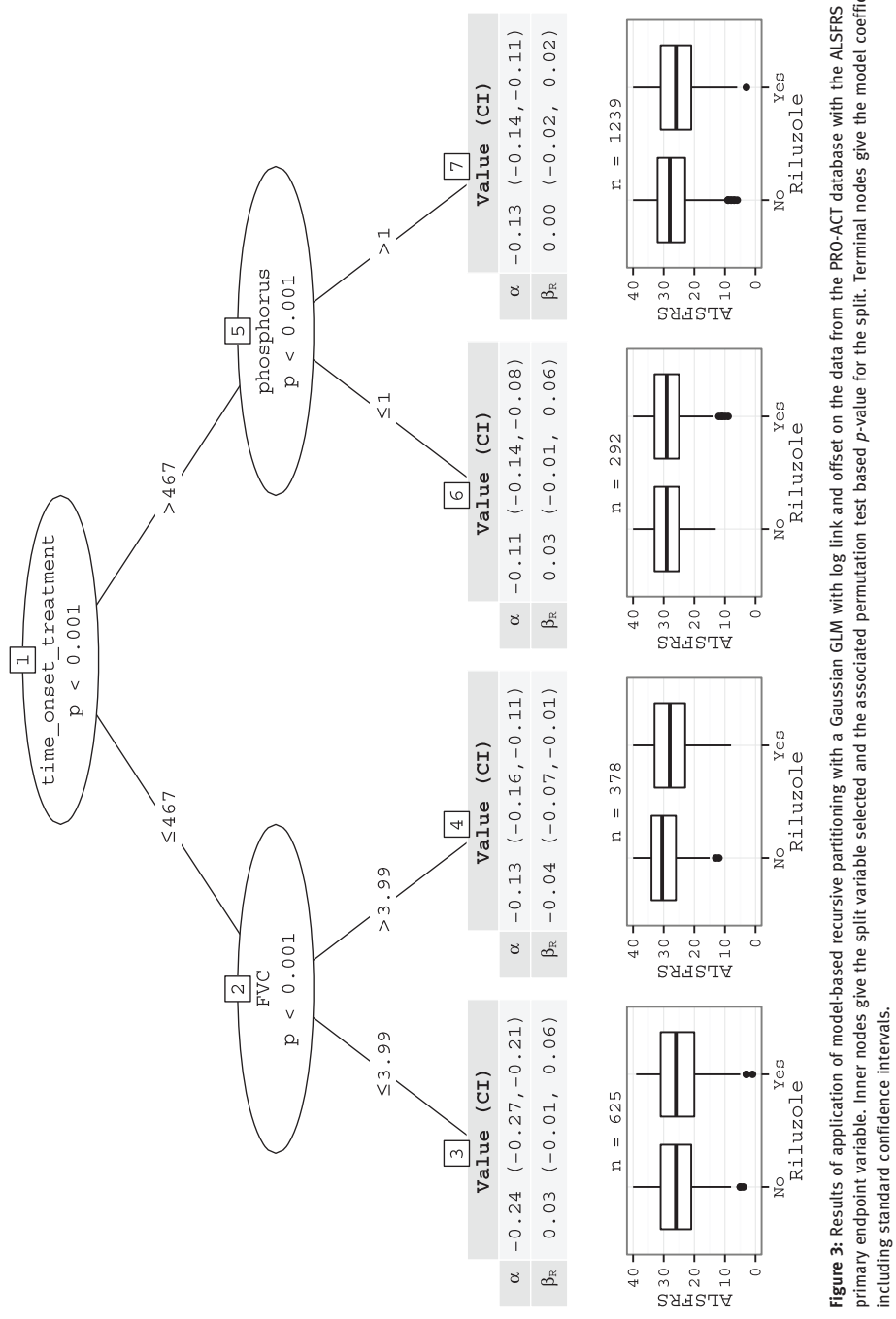


Figure 3: Results of application of model-based recursive partitioning with a Gaussian GLM with log link and offset on the data from the PRO-ACT database with the ALSFRS score as primary endpoint variable. Inner nodes give the split variable selected and the associated permutation test based *p*-value for the split. Terminal nodes give the model coefficients including standard confidence intervals.

3.2 ALSFRS items

The model for the ALSFRS sum-score assumes that the effect of Riluzole is the same for the ten items that define the score. In a more fine-grained analysis, we decomposed the score into its ten items (each ranging between zero and four) and modelled each item by means of a proportional odds model. For one of the ten items assessed at six months, e.g. Y_6, the model reads

$$\mathbb{P}(Y_6 \leq r|\mathbf{X} = \mathbf{x}) = \frac{1}{1 + \exp(-\alpha_r + \beta x_R)}, \tag{14}$$

where $r = 0, \ldots, 4$ is one of the five possible values of Y. The intercept parameters are now $\alpha = (\alpha_0, \ldots, \alpha_3)$ and the partial score function ψ_α is now four dimensional.

As in the previous example, we needed to adjust for the baseline value Y_0, i.e. the value of the ALSFRS item read at the beginning of treatment. This adjustment was implemented by computing separate models; one each for the observations with a start value k, which allows a baseline-specific intercept and treatment effect :

$$\mathbb{P}(Y_6 \leq r|Y_0 = k, X = x) = \frac{1}{1 + \exp(-\alpha_{rk} + \beta_k x_R)} \qquad \text{for } k = 0, \ldots, 4. \tag{15}$$

Therefore, we had a total of five different treatment parameters and 20 different intercepts for each of the ten different items. Model-based recursive partitioning was used to assess the parameter instability of all 250 parameters simultaneously. Note that some of these parameters could not be estimated owing to too small of sample sizes; these were simply discarded.

The implementation of the non-standard model in the theoretical and computational framework of model-based recursive partitioning was straightforward. For every node, we computed the five separate models for the respective baseline values for each of the ten items and extracted the partial scores. A stratified permutation test using the baseline values as independent blocks was used to assess parameter instability. The same procedure was applied for cut-off selection.

The resulting tree (on top of Table 1) contains splits in time between disease onset and treatment start and in the FVC value. The tree is in good agreement with the tree based on the ALSFRS (Figure 3). The third split variable is the lymphocyte percentage. Normal lymphocyte concentrations range from 16 to 33%. Table 1 shows the coefficient values of the models in the terminal nodes for every item and every starting value of the given item. Empty fields indicate that it was not possible to compute the model. Obviously, there were not enough observations in models with zero as starting value for any items in any nodes. The colours in the table indicate whether the effect of Riluzole was positive (blue), negative (pink) or zero (grey). The colours were assigned on the basis of confidence intervals of the coefficient in the given model. Riluzole had a positive effect on patients in the partition of terminal node 3 who had a starting value of 4 in item 1 (speech), 3 (swallowing) or 9 (climbing stairs) and on patients in the partition of terminal node 7 that had a starting value of 3 in item 5 (cutting food and handling utensils). Patients in node 4 who had a starting value of 3 in item 6 (dressing and hygiene) had a negative effect of Riluzole. Riluzole had no effect on patients in the partition of node 6 which are the patients with more than 584 days between disease onset and treatment start who have a lymphocyte concentration under 21.5%.

3.3 Survival time

We used both a Weibull model and a Cox model to identify subgroups with differing effects of Riluzole on the survival endpoint. The application of the model-based recursive partitioning framework in the Weibull model is straightforward and was introduced by Zeileis et al. [7]. Since the Cox model is a semiparametric model, where the intercept is a function of time, treated as a nuisance parameter omitted in the partial likelihood, there is no direct way of obtaining ψ_α. Because, conceptually, deviance residuals are always

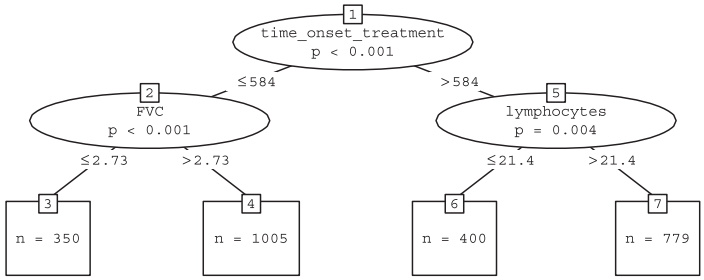


Table 1: Coefficient and confidence interval of Riluzole use in the terminal nodes for every item and every starting value in the model-based recursive partitioning with a proportional odds model (ALSFRS items as outcome). Blue indicates a positive effect of Riluzole, pink a negative effect and grey no effect.

Item	No.	Start	Node 3	Node 4	Node 6	Node 7
Speech	1	0				
		1				
		2		−0.27 (−1.15, 0.60)		
		3	0.33 (−0.33, 1.00)	0.04 (−0.40, 0.47)	−0.27 (−1.12, 0.56)	−0.06 (−0.60, 0.48)
		4	0.84 (0.08, 1.59)	0.11 (−0.28, 0.48)		
Salivation	2	0				
		1				
		2	0.22 (−0.94, 1.40)	0.43 (−0.48, 1.36)	1.36 (−0.31, 3.11)	−0.20 (−1.28, 0.88)
		3	0.15 (−0.57, 0.87)	−0.26 (−0.76, 0.23)	−0.24 (−0.96, 0.47)	−0.05 (−0.58, 0.48)
		4	0.49 (−0.11, 1.07)	−0.03 (−0.39, 0.32)		
Swallowing	3	0				
		1				
		2	0.35 (−0.62, 1.32)	−0.89 (−2.06, 0.21)	1.51 (−0.33, 3.51)	−0.75 (−1.96, 0.43)
		3	0.57 (−0.06, 1.21)	−0.36 (−0.85, 0.12)	−0.40 (−1.17, 0.36)	0.35 (−0.22, 0.93)
		4	0.62 (0.01, 1.23)		0.15 (−0.49, 0.75)	0.28 (−0.22, 0.75)
Handwriting	4	0				−1.45 (−3.21, 0.37)
		1		−1.15 (−2.54, 0.20)	−0.36 (−1.93, 1.25)	−0.08 (−1.26, 1.12)
		2		−0.54 (−1.33, 0.25)		0.04 (−0.71, 0.79)
		3	−0.13 (−0.78, 0.51)	0.14 (−0.22, 0.49)	−0.08 (−0.70, 0.52)	−0.28 (−0.74, 0.17)
		4	−0.10 (−0.71, 0.49)	0.04 (−0.34, 0.42)		−0.14 (−0.65, 0.36)
Cutting	5	0				
		1			−0.01 (−1.19, 1.15)	−0.79 (−1.68, 0.07)
		2		0.15 (−0.54, 0.85)		0.48 (−0.14, 1.12)
		3	−0.03 (−0.76, 0.70)	−0.07 (−0.44, 0.30)	0.10 (−0.60, 0.79)	0.52 (0.03, 1.02)
		4	0.13 (−0.45, 0.72)	−0.09 (−0.49, 0.31)	−0.14 (−0.82, 0.53)	−0.21 (−0.74, 0.30)
Hygiene	6	0				
		1				
		2		−0.11 (−0.65, 0.43)		−0.37 (−0.89, 0.15)
		3	−0.22 (−0.88, 0.44)	−0.37 (−0.72, −0.03)	0.14 (−0.50, 0.78)	0.27 (−0.17, 0.71)
		4	0.26 (−0.40, 0.92)	0.01 (−0.42, 0.44)	0.14 (−0.71, 0.98)	0.30 (−0.31, 0.90)
Bed	7	0				
		1				
		2			−0.03 (−0.96, 0.89)	0.29 (−0.47, 1.04)
		3	0.15 (−0.57, 0.87)	−0.32 (−0.71, 0.08)	−0.12 (−0.74, 0.49)	−0.05 (−0.45, 0.35)
		4	−0.21 (−0.80, 0.36)	−0.10 (−0.45, 0.24)	−0.11 (−0.81, 0.57)	−0.35 (−0.90, 0.18)

(*continued*)

Table 1: (*continued*)

Item	No.	Start	Node 3	Node 4	Node 6	Node 7
Walking	8	0				
		1				
		2			0.48 (−0.22, 1.16)	−0.04 (−0.56, 0.46)
		3		0.11 (−0.33, 0.55)		0.46 (−0.12, 1.04)
		4	0.51 (−0.16, 1.18)	0.13 (−0.27, 0.52)		
Stairs	9	0				
		1		−0.02 (−0.49, 0.45)	−0.01 (−0.72, 0.68)	−0.39 (−0.89, 0.10)
		2			−0.80 (−2.10, 0.48)	0.07 (−0.97, 1.12)
		3		0.26 (−0.19, 0.72)	−0.65 (−1.46, 0.15)	−0.16 (−0.79, 0.46)
		4	1.01 (0.27, 1.77)	0.06 (−0.35, 0.48)	0.72 (−0.11, 1.55)	0.29 (−0.32, 0.89)
Respiratory	10	0				
		1				
		2				
		3				
		4	0.58 (−0.06, 1.24)	−0.08 (−0.44, 0.28)		

defined as the derivative of the log-likelihood with respect to the intercept, we applied martingale residuals as ψ_α. Also worth noting is that both models assume proportional hazards. For the segmented model, proportional hazards are only assumed within each partition. This has to be kept in mind when interpreting the treatment effect in different nodes: Parameters with different signs are clearly linked to opposing treatment effects, but when the parameters only differ in size, it is hard to say whether it is because the groups differ in treatment effect or because they differ in the hazard function.

3.3.1 Weibull model

The Weibull model is a transformation model of the form

$$\mathbb{P}(Y \le y | X = x) = F\left(\frac{\log(y) - \alpha_1 - \beta x_R}{\alpha_2}\right), \tag{16}$$

where F is the cumulative distribution function of the Gompertz distribution. Weibull models are fitted via maximum-likelihood estimation, and therefore the objective function in this case is the negative log-likelihood and the score function has one column per parameter, i.e. intercept α_1, slope parameter β and scale parameter α_2. In the Weibull model, we take the usual intercept as well as the scale parameter as "intercept"-parameter $\alpha = (\alpha_1, \alpha_2)^T$ because they define the shape of the baseline hazard and hence in some respect take the role of an intercept. Splitting in the intercept or scale parameter score suggests non-proportional hazards.

Figure 4 shows that the patient's age and again the time between onset and treatment start play a role in the partitioning. Older patients (> 55.7 years) for whom the time between onset and treatment was longer than 757 days and very young patients did not seem to benefit at all from the treatment. In the remaining two groups, life expectancy seemed to be prolonged for patients treated with Riluzole.

3.3.2 Cox model

The use of the Cox model in model-based recursive partitioning is a rather special case, since the baseline hazard in the Cox model is treated as an infinite-dimensional nuisance parameter and estimation is performed by minimisation of the negative partial log-likelihood. The Cox proportional hazards model is given by

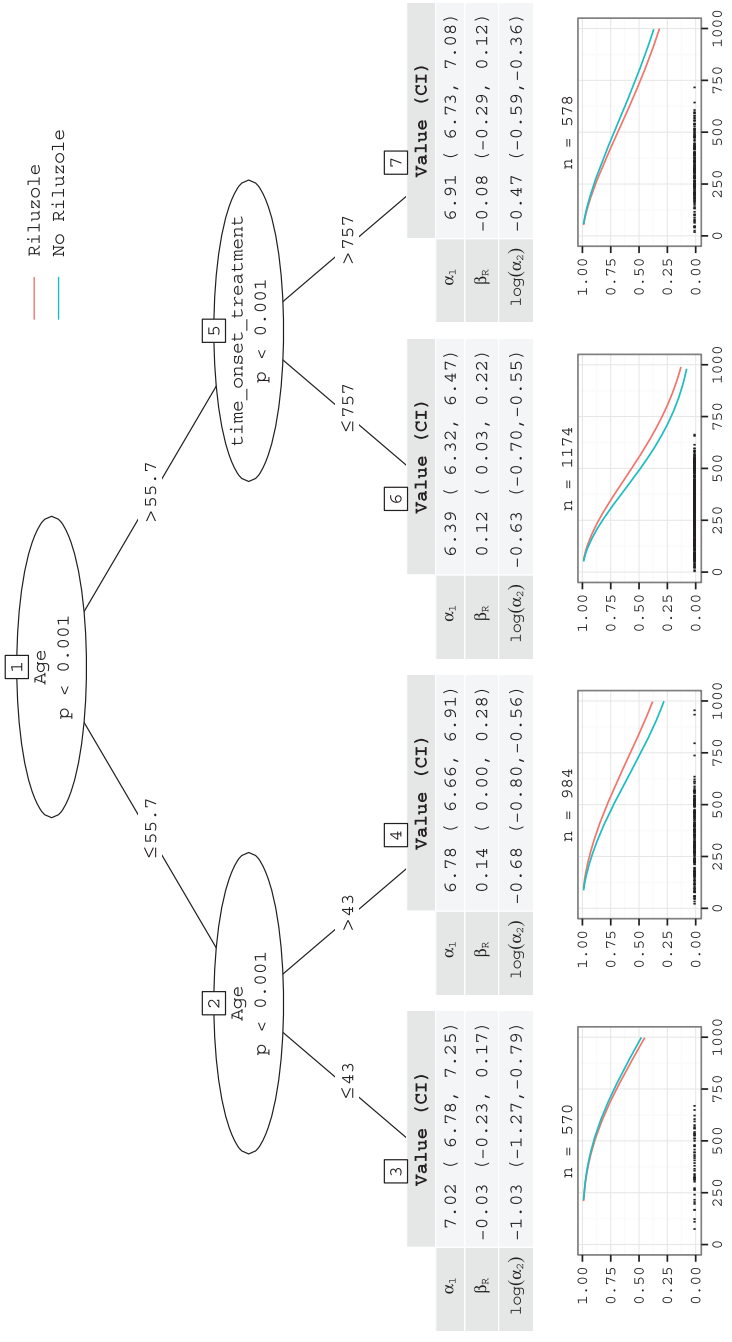


Figure 4: Results of application of model-based recursive partitioning with a Weibull model and data from the PRO-ACT database with survival time as primary endpoint variable. Inner nodes give the split variable selected and the associated permutation test based *p*-value for the split. Terminal nodes give the model coefficients, including standard confidence intervals and the survival curves in the two groups of treatment. Rugs indicate event times.

$$\lambda(y|\mathbf{x}) = \lambda_0(y)\exp(\beta x_R), \tag{17}$$

where λ is the hazard function and λ_0 the baseline hazard function. The partial score function ψ_α (or better, ψ_{λ_0}) cannot be easily derived. As surrogate score function, we propose using the martingale residuals as a score for the baseline hazard, which takes the role of an intercept in the Cox model, and the score residuals for the treatment parameters β. The score residuals are an intuitive choice because they are the first derivative of the partial log-likelihood with respect to the parameters. We used martingale residuals to check whether there is a general difference in the endpoint for different patients, which in parametric models is usually shown by the score with respect to the intercept. Instability in the martingale residuals indicates a violation of the proportional hazards assumption. Since the martingale residuals are not normally distributed, the application of permutation tests is more appropriate than the use of M-fluctuation tests.

Age and the time between disease onset and start of Riluzole treatment form the segments in this example. The tree in this example has almost the same splits as the tree in the previous example. Also estimates support the results of the Weibull example. Again, we did not see much difference between treated and untreated very young patients. For all other groups, Riluzole treatment led to a slight tendency for a lower risk of death (Figure 5).

4 Discussion

Model-based recursive partitioning allows the direct segmentation of the model describing the overall treatment effect as specified in the study protocol. This is the most important benefit of embedding subgroup analysis into this framework because it would be hard to explain why the overall treatment effect and the partitioned treatment effect have to be estimated by two different procedures. This renders the application of suboptimal models unnecessary, such as when a change score is analysed using linear models [21].

Although we are conceptually only interested in finding predictive factors, we think it is necessary to allow splits in the partial scores with respect to both intercept and treatment parameter. This procedure will also detect prognostic factors, but there is a higher chance of including all relevant predictive factors since one might miss prognostic factors when only the treatment scores are split. In our analysis, we decided on the nature of the partitioning variables (prognostic or predictive) only when we interpreted the results of the analysis.

In a model with more covariates than the treatment (e.g. strata), we would still split the partial scores with respect to intercept and treatment parameter for subgroup analyses. A theoretical assumption is then that the parameters that are not split stay constant. In practice, this assumption usually does not hold. It is generally also possible to split more than just the scores with respect to intercept and treatment parameter. Then the split variables are not restricted to being predictive or prognostic but may have an association with the effect of the other covariates.

In model-based recursive partitioning, the variable selection in each node is error controlled, i.e. the probability of selecting a partitioning variable for splitting, when actually all partitioning variables are independent of the scores, is at most as large as the nominal level. The only drawback of using multiple testing procedures is in cases where there are many possible partitioning variables that do not contain information, because with increasing number of noise variables the chance of detecting an actually existing subgroup goes down. The application of permutation tests has the advantage of taking the correlation structure among the partitioning variables into account. Furthermore, for small studies or small subgroups, the exact conditional p-value can be easily approximated up to any desired accuracy; therefore, the method does not rely on asymptotic arguments. The trees obtained by model-based recursive partitioning allow straightforward visualisation, potentially enriched with plots illustrating the distribution of the endpoints

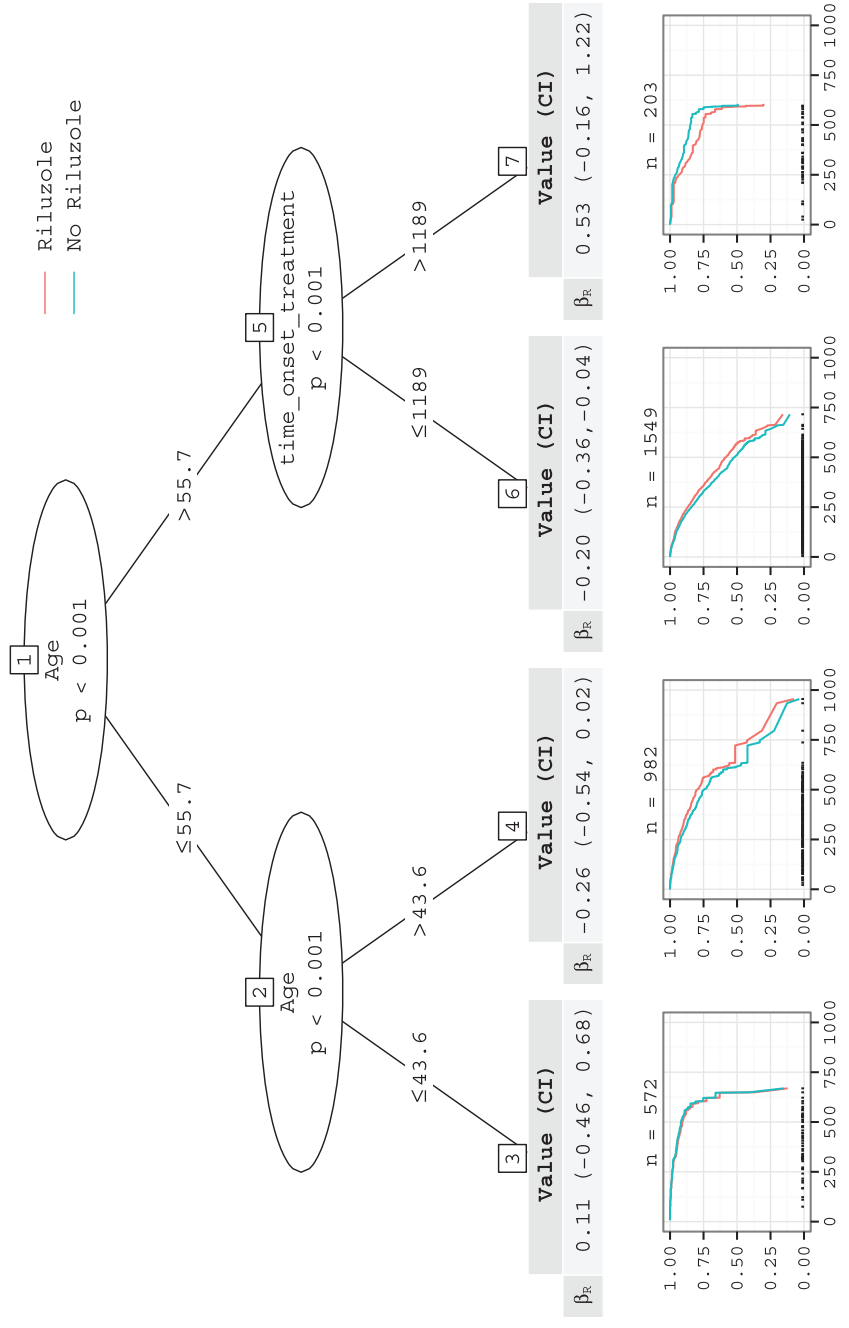


Figure 5: Results of application of model-based recursive partitioning with a Cox model and data from the PRO-ACT database with the survival time as primary endpoint variable. Inner nodes give the split variable selected and the associated permutation test based *p*-value for the split. Terminal nodes give the model coefficients including confidence intervals and the survival curves in the two groups of treatment. Rugs indicate event times.

for the different treatment groups in each subgroup. Therefore, the results of such a subgroup analysis are easily communicated to physicians. Looking at a tree is much easier than trying to understand the meaning of higher-order interactions in a linear predictor. A general drawback of tree methods is the instability of the tree structure with respect to small perturbations in the data, whereas the resulting partitions we are primarily interested in are often relatively stable [13]. Instability in the tree structure can be assessed by means of the variable selection and split statistics, where it is easy to identify all equally likely splits. Bootstrap aggregation and forest procedures are well-known for their ability to stabilise single trees [39] at the cost of interpretability and point into a promising future research direction also for model-based recursive partitioning.

The statistical properties of the confidence intervals derived from the segmented model await further attention. Leeb and Pötscher ([40]) discuss the validity of inference after variable selection and claim that it is difficult if at all possible. Bai and Perron [41], who discuss the construction of confidence intervals after splitting up the data based on a break point in a single partitioning variable, argue that it is possible. In our approach we first search for the most appropriate partitioning variable (variable selection) and then search for the optimal split point (break point selection). To our knowledge there is no literature on inference after variable and break point selection and thus it is unclear if or how valid confidence intervals can be computed. In any case the results of such a subgroup analysis have to be confirmed in follow-up trials, which lowers the necessity of confidence intervals. To be conservative one can see the confidence intervals for parameters in the subgroup-specific models as shown in our examples as a range of possible values and hence as a measure of variability rather than significance [42].

It would be interesting to extend the framework of the PRO-ACT database of ALS studies to models for non-independent data, such as mixed models for longitudinal observations. This would allow ALS disease progression to be modelled over time, and also a potentially time-varying treatment effect to be assessed. In our way of modelling the functional endpoint, we include no information about patients that died within the first six months after treatment start. Joint modelling of the longitudinal functional endpoint and the survival endpoint is a means of combining all possible information [43].

Despite the deficits of model-based recursive partitioning for subgroup analysis discussed in this section, we think that the procedure as introduced and illustrated in this paper rather closely resembles the requirements for statistical procedures in this field as outlined in the EMA guideline [1]. In particular, it is the most generally applicable procedure with statistical error control and unbiased variable selection [13, 7]. With the available open-source implementation (see following section for details), the method can be applied straightforwardly elsewhere.

Computational details

An open-source implementation of all methods discussed in this paper and beyond is available in the **partykit** package hothorn_partykit:_2014. PRO-ACT data are available at `https://nctu.partners.org/ProACT/` [44]. The source code for reading and cleaning the database is provided in the **TH.data** package [45]. The source code for the analyses is provided in the supplementary material. All computations were conducted using **partykit** (version 0.8-2) in the R system for statistical computing [46], version 3.1.2).

Listing 1: Code snippet for Weibull model in model-based recursive partitioning using the function `ctree()` from the **partykit** package.

```
## Function to compute Weibull model and return score matrix
mywb <- function(data, weights, parm) {
  mod <- survreg(Surv(survival.time, cens) Riluzole,
                 data = data, subset = weights > 0,
                 dist = "weibull")
```

```
  ef <- as.matrix(estfun(mod)[,parm])
  ret <- matrix(0, nrow = nrow(data), ncol = ncol(ef))
  ret[weights > 0,] <- ef
  ret
}
## Compute tree
tree <- ctree(fm, data = data, ytrafo = my.wb,
              control = ctree_control(maxdepth = 2,
                                      testtype = "Bonferroni"))
```

Acknowledgements: We are thankful to the organisers and participants of the "Workshop on Classification and Regression Trees" (March 2014), sponsored by the Institute for Mathematical Sciences of the National University of Singapore, for helpful feedback and stimulating discussions and to Karen A. Brune for improving the language. Financial support by the Swiss National Science Foundation (grant 205321_163456) is gratefully acknowledged.

References

1. European Medicines Agency. EMA guideline on the investigation of subgroups in confirmatory clinical trials (draft). Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/02/WC500160523.pdf, 2014.
2. Gadbury GL, Iyer HK. UnitTreatment Interaction and Its Practical Consequences. Biometrics 2000;56:882–5.
3. Holland PW. Statistics and causal inference. J Am Stat Assoc 1986;81:945–60.
4. Morgan JN, Sonquist JA. Problems in the analysis of survey data, and a proposal. J Am Stat Assoc 1963;58:415–34.
5. Doove LL, Dusseldorp E, Van Deun K, Van Mechelen I. A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment-subgroup interactions. Adv Data Anal Class 2014;8:403–25.
6. Italiano A. Prognostic or predictive? It's time to get back to definitions!. J Clin Oncol 2011;29:4718–4718.
7. Zeileis A, Hothorn T, Hornik K. Model-based recursive partitioning. J Comput Graph Stat 2008;17:492–514.
8. Hastie T, Tibshirani R. Varying-Coefficient Models. J R Stat Soc Series B (Methodological) 1993;55:757–96. Available at: http://www.jstor.org/stable/2345993.
9. Zeileis A, Hornik K. Generalized M-fluctuation tests for parameter instability. Stat Neerl 2007;61:488–508.
10. Zeileis A, Hothorn T. A toolbox of permutation tests for structural change. Stat Papers 2013;54:931–54.
11. Hothorn T, Hornik K, Van de Wiel MA, Zeileis A. A Lego system for conditional inference. Am Stat 2006a;60:257–63.
12. Hothorn T, Hornik K, van de Wiel MA, Zeileis A. Implementing a class of permutation tests: The coin package. J Stat Software 2008;28:1–23.
13. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: A conditional inference framework. J Comput Graph Stat 2006b;15:651–74.
14. Loh W-Y, He X, Man M. A regression tree approach to identifying subgroups with differential treatment effects. Stat Med 2015;34:1818–33.
15. Foster JC, Taylor JM, Ruberg SJ. Subgroup identification from randomized clinical trial data. Stat Med 2011;30:2867–80.
16. Kehl V, Ulm K. Responder identification in clinical trials with censored data. Comput Stat Data Anal 2006;50:1338–55.
17. Ciampi A, Negassa A, Lou Z. Tree-structured prediction for censored survival-data and the Cox model. J Clin Epidemiol 1995;48:675–89.
18. Negassa A, Ciampi A, Abrahamowicz M, Shapiro S, Boivin JF. Tree-structured subgroup analysis for censored survival data: Validation of computationally inexpensive model selection criteria. Stat Comput 2005;15:231–9.
19. Loh W-Y. Regression trees with unbiased variable selection and interaction detection. Stat Sinica 2002;12:361–86.
20. Loh W-Y. Improving the precision of classification trees. Ann Appl Stat 2009;3:1710–37.
21. Dusseldorp E, Van Mechelen I. Qualitative interaction trees: A tool to identify qualitative treatment-subgroup interactions. Stat Med 2013;33:219–37.
22. Dusseldorp E, Doove L, Van Mechelen I. quint: Qualitative Interaction Trees. Available at: http://CRAN.R-project.org/package=quint, R package version 1.0., 2013
23. Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect search – A recursive partitioning method for establishing response to treatment in patient subpopulations. Stat Med 2011;30:2601–21.

24. Lipkovich I, tri enko AD. Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using SIDES. J Biopharm Stat 2014;24:130–53.
25. Gama J. Functional trees. Mach Learn 2004;55:219–50.
26. Quinlan JR. C4.5: Programs for Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 1993.
27. Kim H, Loh W-Y. Classification trees with unbiased multiway splits. J Am Stat Assoc 2001;96:589–604.
28. Chan K-Y, Loh W-Y. Lotus: An algorithm for building accurate and comprehensible logistic regression trees. J Comput Graph Stat 2004;13:826–52.
29. Su X, Wang M, Fan J. Maximum likelihood regression trees. J Comput Graph Stat 2004;13:586–98.
30. Chipman H, George E, McCulloch R. Bayesian treed models. Mach Learn 2002;48:299–320.
31. Bernardo JM, Bayarri MJ, Berger JO, Dawid AP, Heckerman D, Smith AFM, et al. Bayesian Treed Generalized Linear Models. Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting, 85, 2003.
32. European Medicines Agency. Riluzole Zentiva: EPAR summary for the public. Available at:http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Summary_for_the_public/human/002622/WC500127609.pdf, 2012.
33. Chiò A, Logroscino G, Hardiman O, Swingler R, Mitchell D, Beghi E, et al., and On Behalf of the Eurals Consortium. Prognostic factors in ALS: A critical review. Amyotroph Lateral Scler 2009;10:310–23.
34. Atassi N, Berry J, Shui A, Zach N, Sherman A, Sinani E, et al. The PROACT database: Design, initial analyses, and predictive features. Neurology 2014;83:1719–25.
35. Küffner R, Zach N, Norel R, Hawe J, Schoenfeld D, Wang L, et al. Crowd sourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. Nat Biotechnol 2014;33:51–7.
36. Brooks BR, Sanjak M, Ringel S, England J, Brinkmann J, Pestronk A, et al. The amyotrophic lateral sclerosis functional rating scale – assessment of activities of daily living in patients with amyotrophic lateral sclerosis. Arch Neurol 1996;53:141–7.
37. Franchignoni F, Mora G, Giordano A, Volanti P, Chio A. Evidence of multidimensionality in the ALSFRS-R scale: A critical appraisal on its measurement properties using Rasch analysis. J Neurol Neurosurg Psychiatry 2013;84:1340–5.
38. Hothorn T, Jung HH. RandomForest4life: A Random Forest for predicting ALS disease progression. Amyotroph Lateral Scler Frontotemporal Degener 2014;15:444–52.
39. Strobl C, Malley J, Tutz G. An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. Psychol Methods 2009;14:323–48.
40. Leeb H, Pötscher BM. Model selection and inference: Facts and fiction. Economet Theory 2005;21:21–59.
41. Bai J, Perron P. Computation and analysis of multiple structural change models. J Appl Economet 2003;18:1–22.
42. Lagakos SW. The challenge of subgroup analyses–reporting without distorting. N Eng J Med 2006;354:1667–9.
43. Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. Biostatistics 2000;1:465–80.
44. Massachusetts General Hospital, N. C. R. I. Pooled resource open-access ALS clinical trials database. Available at: https://nctu.partners.org/ProACT/, 2013.
45. Hothorn T. TH.data: TH's Data Archive. Available at: http://CRAN. R-project.org/package=TH.data, R package version 1.0-4, 2014.
46. R Core Team. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org/ 2014.
47. Hothorn T, Zeileis A. partykit: A modular toolkit for recursive partitioning in R, Technical report. Available at: http://eeecon.uibk.ac.at/wopec2/repec/inn/wpaper/2014-10.pdf, 2014.

PAPER II

Individual treatment effect prediction for amyotrophic lateral sclerosis patients

Heidi Seibold, Achim Zeileis, Torsten Hothorn

Published in *Statistical Methods in Medical Research*, 2017, online first.

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–22

DOI: 10.1177/0962280217693034
journals.sagepub.com/home/smm
SAGE

Individual treatment effect prediction for amyotrophic lateral sclerosis patients

Heidi Seibold,[1] Achim Zeileis[2] and Torsten Hothorn[1]

Abstract

A treatment for a complicated disease might be helpful for some but not all patients, which makes predicting the treatment effect for new patients important yet challenging. Here we develop a method for predicting the treatment effect based on patient characteristics and use it for predicting the effect of the only drug (Riluzole) approved for treating amyotrophic lateral sclerosis. Our proposed method of model-based random forests detects similarities in the treatment effect among patients and on this basis computes personalised models for new patients. The entire procedure focuses on a base model, which usually contains the treatment indicator as a single covariate and takes the survival time or a health or treatment success measurement as primary outcome. This base model is used both to grow the model-based trees within the forest, in which the patient characteristics that interact with the treatment are split variables, and to compute the personalised models, in which the similarity measurements enter as weights. We applied the personalised models using data from several clinical trials for amyotrophic lateral sclerosis from the Pooled Resource Open–Access Clinical Trials database. Our results indicate that some amyotrophic lateral sclerosis patients benefit more from the drug Riluzole than others. Our method allows gradually shifting from stratified medicine to personalised medicine and can also be used in assessing the treatment effect for other diseases studied in a clinical trial.

Keywords

Personalised medicine, individual treatment effect, random forest, model-based recursive partitioning

1 Introduction

Amyotrophic lateral sclerosis (ALS) is a deadly disease that affects motor neurons in the brain and spinal cord, i.e. the neurons responsible for voluntary muscle control. Riluzole (Rilutek) is the only approved drug for this disease to date. According to the European Medicines Agency,[1] Riluzole prolongs the median survival of ALS patients, depending on the dose, by a few months. Several side effects, such as sickness, weakness or increased liver enzyme levels are mentioned.[1] Knowledge how Riluzole works on the nervous system of ALS patients is limited. The Pooled Resource Open–Access Clinical Trials (PRO-ACT) database[2] is the largest database containing clinical trial data of ALS patients available and was initiated to retrieve more information on the disease. It contains data from 17 ALS studies conducted between 1990 and 2010. Using these data, we aimed at finding out more about the effect of Riluzole on the health and survival of patients.

Before statistical analyses and *p*-values entered into medical progress 70 years ago, doctors treated patients individually based on their experiences and knowledge.[3] Since the beginning of the 'golden age of randomised clinical trials', however, medication became more and more standardised. Nowadays, much knowledge about the effect of drugs has accumulated, cornerstone drugs such as antibiotics have been used for decades and many diseases can be treated successfully; however, providing new drugs for the general public becomes more difficult. Diseases such as ALS are too complex to treat all patients in the same way. Therefore, there is a need to return to more individualised treatments, but this time with the use of statistical concepts.

[1]Department of Biostatistics, Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Switzerland
[2]Department of Statistics, Faculty of Economics and Statistics, University of Innsbruck, Austria

Corresponding author:
Torsten Hothorn, Biostatistics Department, Epidemiology, Biostatistics & Prevention Institute, University of Zurich, Hirschengraben 84, CH-8001 Zürich, Switzerland.
Email: Torsten.Hothorn@uzh.ch

In the past years, there has been an immense effort towards personalised medicine in the analysis of randomised controlled trials. The goal is to identify predictive factors, i.e. factors that interact with the treatment,[4] such as biomarkers, other treatments and environmental circumstances. In the following, we will refer to these factors as patient characteristics. Prognostic factors, i.e. factors that directly affect the patient's outcome, are only of secondary interest, but should not be neglected, because they not only change the general level of the outcome – showing in the individual intercept – but might also be predictive and prognostic.[5] Note that we use the terms predictive and prognostic as in the medical literature,[4] but in a statistical sense both groups of variables are useful predictors. For drugs for which the biological mode of action is unknown, predictive and prognostic factors should first be identified in a data-driven way. New hypotheses can then be generated and new trials can be planned based on these hypotheses. In this first step, we ask *whether* a certain patient characteristic is relevant and not *why*.

Many new statistical methods in the field of stratified medicine, i.e. subgroup analysis, have been developed. Subgroup analyses aim at finding groups of patients that have differential treatment effects. Most of the methods are based on recursive partitioning (trees) and/or interaction models.[6–12] The tree-based methods for subgroup analyses have specialised splitting procedures for partitioning the patients into groups with higher and lower treatment effect. Interaction models evaluate the interaction between the treatment and given patient characteristics. The idea behind methods of subgroup analyses in general is to obtain a treatment effect $\beta(\mathbf{z})$ that depends on the patient characteristics $\mathbf{z}$. For example, the treatment effect could depend on the age of patients, in which patients less than 40 years of age improve through the treatment, patients between 40 and 60 do not improve and patients older than 60 years improve, but less than the patients under 40 years:

$$\beta(\mathbf{z}) = \begin{cases} 1 & \text{if } z_{\text{age}} < 40 \\ 0 & \text{if } 40 \leq z_{\text{age}} < 60 \\ 0.5 & \text{if } 60 \leq z_{\text{age}} \end{cases} \tag{1}$$

However, the assumption that the treatment effect is a step function may be too restrictive, and $\beta(\mathbf{z})$ in reality may be a smooth interaction function. In other words, personalised medicine is required instead of stratified medicine. Because methods for subgroup analyses again generalise the treatment effect for a group of patients, it can only be considered as a step in the direction toward personalised medicine. We provide a method that can estimate smooth treatment effect functions using model-based random forests and weighted models. More importantly, this method provides an estimate for the treatment effect of a future patient, thereby allowing a decision to be made whether treatment of this patient is appropriate.

2 Methods

Seibold et al., 2016[5] introduced a means of conducting subgroup analysis for randomised controlled trials using model-based recursive partitioning. One first defines a model $\mathcal{M}((Y, \mathbf{X}), \vartheta)$ with primary endpoint Y, covariates $\mathbf{X}$ including the randomised treatment indicator

$$X_A = \begin{cases} 1 & \text{if patient received the (new) treatment} \\ 0 & \text{if patient received no treatment (or standard of care),} \end{cases} \tag{2}$$

and parameter vector ϑ. In the following, we will consider likelihood models (e.g. generalised linear models or parametric survival models) where the model parameters ϑ can be estimated by maximising the log-likelihood $\ell((Y, \mathbf{X}), \vartheta)$ of those models (e.g. Gaussian log-likelihood or Weibull log-likelihood) or equivalently by solving the score equation

$$\sum_{i=1}^{N} s((y, \mathbf{x})_i, \vartheta) = \mathbf{0} \tag{3}$$

with

$$s((y, \mathbf{x})_i, \vartheta) = \frac{\partial \ell((y, \mathbf{x})_i, \vartheta)}{\partial \vartheta} \tag{4}$$

In most applications, the model contains only an intercept α and a treatment effect β, i.e. $\mathbf{X} = (1, X_A)$ and $\vartheta = (\alpha, \beta)^\top$, but more parameters are possible, such as coefficients of additional regressors or scale and shape parameters for the response distribution. Technically, there can also be more than two treatment groups or no intercept. For simplicity, we will focus on the basic case with intercept and treatment effect and two treatment groups. The method obtains subgroups $\{\mathcal{B}_{b=1,\ldots,B}\}$ that differ with regard to the treatment effect β and potentially the intercept α. The subgroups are defined by patient characteristics $\mathbf{Z} = (Z_1, \ldots, Z_J) \in \mathcal{Z}$. Hence, the intercept and treatment parameters can be written as a function of the subgroup-defining variables $\mathbf{z}$. In other words, the patient characteristics $\mathbf{Z}$ are not part of the model $\mathcal{M}((Y, \mathbf{X}), \vartheta)$ but are used to define the subgroups in which the model parameters differ, and then the model parameters are estimated within each subgroup.

Conceptually, the partitioned model parameters $\alpha(\mathbf{z})$ and $\beta(\mathbf{z})$ might depend on $\mathbf{z}$ in a more complex way than a simple tree structure. Therefore, the model parameters are not step functions, but rather smooth interaction functions, so that an individual treatment effect (as in personalised medicine) can be computed for each patient instead of only for each subgroup of patients (as in stratified medicine). The function $\beta(\mathbf{z})$ can then be understood as an estimate of the counterfactual individual treatment effect of a patient with patient characteristics $\mathbf{z}$.

The most intuitive step from a tree structure to a more complex structure is to use a random forest instead of a single tree. Hence, we propose a strategy in which a model-based random forest is used to measure how similar patients are with respect to the treatment effect and the treatment effect of each patient is predicted on this basis using personalised models.

2.1 Random forest

Random forests[13] compute an ensemble of T trees. The proposed algorithm draws subsamples $\mathcal{L}_t$, $t = 1, \ldots, T$ of the given N observations and fits a model-based tree to each subsample using a randomly sampled set of candidate split variables $\mathbf{z}$. The data $\mathcal{L}_t^c$ that were not in the learning sample for tree t are called out-of-bag data. Classical random forests provide information on the similarity between observations with respect to the response. Model-based random forests provide information on the similarity between observations (patients) with respect to the model parameters, i.e. treatment effect and intercept.

This section focuses on the estimation of the trees, and the following section features the computation of the similarity measure and how the forest can be used to estimate personalised treatment effects.

2.2 Split procedure

The special feature of our method is the split procedure, which is based on the empirical estimating function

$$\boldsymbol{s} = \begin{pmatrix} s_\alpha((y, \mathbf{x})_1, \widehat{\vartheta}) & s_\beta((y, \mathbf{x})_1, \widehat{\vartheta}) \\ s_\alpha((y, \mathbf{x})_2, \widehat{\vartheta}) & s_\beta((y, \mathbf{x})_2, \widehat{\vartheta}) \\ \vdots & \vdots \\ s_\alpha((y, \mathbf{x})_N, \widehat{\vartheta}) & s_\beta((y, \mathbf{x})_N, \widehat{\vartheta}) \end{pmatrix} \tag{5}$$

which contains the score contributions $s_\alpha((y, \mathbf{x})_i, \widehat{\vartheta})$ and $s_\beta((y, \mathbf{x})_i, \widehat{\vartheta})$. The score contributions are the partial derivatives of the log-likelihood with respect to α or β, respectively, evaluated at the N observed data points and the estimated parameters $\widehat{\vartheta} = (\hat{\alpha}, \hat{\beta})^\top$.[14] The matrix of score contributions $\boldsymbol{s}$ contains information on the deviation from the model fit for all parameters and observations of a given model $\mathcal{M}((Y, \mathbf{X}), \vartheta)$. The contributions can thus be seen as residuals. Score contributions are widely used in model inference (e.g. see Chapter 3.7, Tutz, 2012)[15] and in recursive partitioning.[14,16] They are particularly useful because they fluctuate randomly around 0 in well-fitting models, and they show patterns when there are parameter instabilities.

To obtain a split in model-based recursive partitioning for this setup, the following steps have to be performed:

- Estimate the parameters in the prespecified model $\mathcal{M}((Y, \mathbf{X}), \vartheta)$.
- Compute the associated score matrix $\boldsymbol{s}$.

- Perform tests of independence between the score contributions and the partitioning variables:

$$H_0^{\alpha,j}: \quad s_\alpha((Y,\mathbf{X}),\widehat{\vartheta}) \quad \perp \quad Z_j$$
$$H_0^{\beta,j}: \quad s_\beta((Y,\mathbf{X}),\widehat{\vartheta}) \quad \perp \quad Z_j \quad j=1,\ldots,J$$

  The smallest p-value corresponds to the greatest deviation from the model assumption; that intercept and treatment parameter are the same for all patients in the given node/subgroup.
- If any p-value is lower than the significance level, select the partitioning variable that has the highest association (lowest p-value) to any of the relevant residuals for the split.
- Search for the optimal split point in the selected partitioning variable using a suitable criterion, such that the models in the resulting daughter nodes have as little association between the partitioning variable and the residuals as possible.

This split procedure is repeated until a stopping criterion is met. This can be, for example, when no p-values are lower than the significance level or if subgroups become too small. For detailed information on stopping criteria, see Hothorn et al., 2015.[17] In the end, a tree is obtained with disjoint subgroups

$$\dot{\bigcup_b} \mathcal{B}_b = \mathcal{Z} \tag{6}$$

Accordingly in a random forest of T trees, each tree defines disjoint subgroups

$$\dot{\bigcup_b} \mathcal{B}_{tb} = \mathcal{Z} \quad \forall t = 1,\ldots,T \tag{7}$$

The independence tests can be performed using permutation tests[18,19] or, for reasonably large samples, using M-fluctuation tests.[14,20] Unbiased recursive partitioning methods commonly use tests with node-wise null hypotheses of 'no further split needed', as we do here.[14,16,19] Since one test is computed per patient characteristic eligible in the given node, multiplicity adjustment such as Bonferroni correction is recommended. More details on the algorithm and the test procedures used are documented in Appendix 2.

2.3 Personalised models

In personalised medicine, the goal is to learn how much a person will profit from a given treatment and what would happen if the standard of care or no treatment is given. For any patient, it is possible to compute a personalised model based on the similarity of this observation to the observations in the training data. In general, any measure of similarity $w_i(\mathbf{z}_k)$ between patients i and k with respect to the treatment effect and general health could be used, i.e. any measure that compares patients i and k in terms of $\beta(\mathbf{z}_i)$ to $\beta(\mathbf{z}_k)$ and of $\alpha(\mathbf{z}_i)$ to $\alpha(\mathbf{z}_k)$. A straight forward similarity measure in this sense is the number of times patients i and k are classified in the same subgroup by the single model-based trees in the random forest

$$w_i(\mathbf{z}_k) = \sum_{t=1}^{T} \sum_{b=1}^{B_t} (\mathbf{z}_i \in \mathcal{B}_{tb}) \wedge (\mathbf{z}_k \in \mathcal{B}_{tb}) \tag{8}$$

with T being the number of trees used for the computation of the forest and B_t being the number of subgroups from tree t.[21–23] If patient i is part of the training set, the weights can be computed out-of-bag, i.e. the only trees ($t = 1,\ldots,T$) considered are those where patient i is not in the subset $\mathcal{L}_t$ for the computation.

To obtain the personalised model $\mathcal{M}((Y,\mathbf{X}),\widehat{\vartheta}(\mathbf{z}_i))$ for patient i, the base model is recomputed with the weighted training data, which is equivalent to minimising the personal log-likelihood of patient i (the sum of weighted log-likelihood contributions)

$$\underset{\vartheta}{\operatorname{argmax}} \sum_{k=1}^{N} w_i(\mathbf{z}_k) \cdot \ell\left((y,\mathbf{x})_k, \vartheta(\mathbf{z}_i)\right) \tag{9}$$

In other words, every patient k from the training set is included $w_i(\mathbf{z}_k)$ times in the 'new data set' to compute the personalised model for patient i. In the following, the parameters estimated from this model will be denoted by $\widehat{\vartheta}(\mathbf{z}_i) = (\hat{\alpha}(\mathbf{z}_i), \hat{\beta}(\mathbf{z}_i))$.

Using the personalised models, it is possible to obtain a log-likelihood. From the personalised model for patient i, the log-likelihood contribution $\ell((y, \mathbf{x})_i, \widehat{\vartheta}(\mathbf{z}_i))$ for this observation is computed. The log-likelihood then is

$$\sum_{i=1}^{N} \ell\Big((y, \mathbf{x})_i, \widehat{\vartheta}(\mathbf{z}_i)\Big) \tag{10}$$

which we refer to as forest log-likelihood. A variant of this algorithm for non-personalised transformation models is discussed in Hothorn and Zeileis.[24]

2.4 Improvement through personalised models

To check whether the personalised models actually lead to an improvement of the base model, one tests the hypothesis

$$H_0 : \tag{11}$$

$$\underbrace{\alpha(\mathbf{Z}) \equiv \alpha}_{H_0^\alpha} \tag{12}$$

$$\bigcap$$

$$\underbrace{\beta(\mathbf{Z}) \equiv \beta}_{H_0^\beta}$$

This strict null hypothesis is to be rejected if any of the patient characteristics contain information on the outcome or the treatment effect. To conduct the test, one can proceed as follows:

- Compute the forest log-likelihood and the log-likelihood of the base model and calculate their difference. This difference is a measure of how much better the personalised models are compared to the base model.
- Draw parametric bootstrap samples from the base model.
- Compute the forest log-likelihood and the log-likelihood of the base model in the bootstrap samples and again compute the differences. The distribution of these values represents the distribution under the null hypothesis.
- The p-value is then the proportion of bootstrap samples in which the difference in log-likelihoods exceeds the observed difference in the original data. Note, that this p-value will be very low or even 0 when the patient characteristics contain information on the outcome or the treatment effect.

In practice, one may be interested in just H_0^β, but testing the sub-hypotheses H_0^α and H_0^β separately is not straight-forward. An approximation would be to compute the personalised models using a forest that splits based only on the partial score function with respect to α or β. Patient characteristics, however, are often not exclusively predictive or prognostic but can be both. Also, if a patient characteristic is purely prognostic, this still may result in a pattern in both partial score functions. For more details, see Seibold et al, 2016.[5]

2.5 Dependence plots

A partial dependence plot describes the dependence of a function (in our case the treatment effect $\hat{\beta}(\mathbf{z})$) and a variable (in our case, a partitioning variable).[25] The partial dependence plot resulting from a model-based tree would show a step function. The partial dependence from a random forest can be smoother for continuous partitioning variables. It can be obtained by plotting $\hat{\beta}(z_j)$ against z_j for each partitioning variable $j = 1, \ldots, J$.

2.6 Variable importance

The variable importance for the random forest is computed based on the tree log-likelihoods. For a given forest computed with T trees, the log-likelihood is computed as follows:

- Select the out-of-bag data $\mathcal{L}_t^c$ and determine the terminal node/subgroup to which each observation i belongs.

- Compute the log-likelihood contribution of each observation $i \in \mathcal{L}_t^c$ based on the respective model in the terminal node/subgroup with parameters $\widehat{\vartheta}(\mathbf{z}_i)$.
- Compute the out-of-bag log-likelihood as the sum of the contributions

$$\ell_t = \sum_{i \in \mathcal{L}_t^c} \ell\left((y, \mathbf{x})_i, \widehat{\vartheta}(\mathbf{z}_i)\right) \tag{13}$$

To obtain the variable importance of a given variable z_j, $j = 1, \ldots, J$, the variable is permuted. The log-likelihood is computed as above, except that the column with information about z_j in the out-of-bag data is replaced by the permuted z_j. We denote the log-likelihood of tree t with variable z_j permuted by $\ell_t^{(j)}$. The variable importance is then

$$\mathrm{VI}_j = \frac{1}{T}\sum_{t=1}^{T}\left[\ell_t - \ell_t^{(j)}\right] \tag{14}$$

If the variable importance is high, the variable is an important predictive and/or prognostic factor. Note that due to the signed differences, the variable importances might become negative signalling that the log-likelihood merely improved by chance and that the variable is not important. As the size of the negative values conveys information on the overall importance variability, we do not collapse to 0 which would otherwise be a sensible restriction. It is possible to compute also conditional variable importances[26] to account for correlation between patient characteristics. In the following, we focus on unconditional variable importances.

3 Results

3.1 PRO-ACT data

The PRO-ACT (https://nctu.partners.org/ProACT) database contains longitudinal data of ALS patients that participated in one of 16 phase II and III trials and one observational study. It is a project initiated by the non-profit organisation Prize4Life (http://www.prize4life.org/) to enhance knowledge about ALS. It contains information on a broad variety of patient characteristics, such as vital signs, the patient's and family's history, and treatment information. Identification criteria, such as study centres, are not included in the database. Also collected are the survival time and the ALS functional rating scale (ALSFRS), which is a score measuring the patients' ability of living a normal life.[27] The ALSFRS is a sum-score of 10 items, each of which ranges between 0 and 4, where 0 represents complete inability and 4 represents normal ability. The items are speech, salivation, swallowing, hand-writing, cutting food and handling utensils, dressing and hygiene, turning in bed and adjusting bed clothes, walking, climbing stairs and breathing. As outcomes in the study, we used both the survival time (denoted by survival) and the ALSFRS 6 months after treatment start (denoted by ALSFRS_6) and identified patient characteristics that influence the effect of Riluzole on these outcomes. For the two outcome variables, we obtained two different data sets. We only included observations that contain information on the respective outcome variable and only patient characteristics that have fewer than 50% missing values. The survival time data set contains 3306 observations and 18 patient characteristics. The ALSFRS data set contains 2534 observations and 57 patient characteristics.

Tables 1 and 2 show the estimates including standard errors obtained from the base model for each outcome. For the ALSFRS, this base model is given by

$$\mathbb{E}\left(\frac{\mathrm{ALSFRS}_6}{\mathrm{ALSFRS}_0}\middle| X = x\right) = \frac{\mathbb{E}(\mathrm{ALSFRS}_6 | X = x)}{\mathrm{ALSFRS}_0} = \exp\{\alpha + \beta x_A\} \tag{15}$$

Table 1. ALSFRS base model (Gaussian generalised linear model with log-link and offset).

	Estimate	Std. error	2.5%	97.5%
α	−0.1595	0.0065	−0.1722	−0.1468
β	0.0091	0.0077	−0.0060	0.0242

Given are the parameter estimates, their standard error and the Wald confidence interval.

which represents a Gaussian generalised linear model with log-link and offset log(ALSFRS$_0$), where ALSFRS$_0$ is the ALSFRS that was measured at the time of treatment start. The base model for the survival time is given by the Weibull model

$$\mathbb{P}(T \leq \text{survival} | X = x) = F\left(\frac{\log(\text{survival}) - \alpha_1 - \beta x_A}{\alpha_2}\right) \tag{16}$$

where F is the cumulative distribution function of the Gompertz distribution. Note that the Weibull model has a scale parameter in addition to the intercept, so that both α_1 and α_2 control the appearance of the baseline hazard. In the notation of equation (4), this leads to $\vartheta = (\alpha_1, \alpha_2, \beta)^\top$.

3.2 Personalised models

We computed personalised models for all observations in the respective training data, which were used to obtain the random forest. The distribution of parameter estimates in the personalised models is given in Figure 1 for the ALSFRS and in Figure 2 for the survival time. Figure 1 shows that all patients are predicted to have a positive Riluzole effect, i.e. for all patients taking Riluzole, a higher ALSFRS is achieved compared to those not taking Riluzole. However, there is a variability in the treatment effects, and the distribution of the treatment effect is bimodal (as is the distribution of the intercept). The treatment effect estimated from the base model is between the two modes. The lowest treatment effect a person in this data set is predicted to have is 0.0027.

For the survival time, the lowest predicted treatment effect is 0.0717. However, the value of the treatment effect in the personalised survival models cannot be interpreted in isolation; its meaning depends on the shape of the

Table 2. Survival time base model (Weibull model).

	Estimate	Std. error	2.5%	97.5%
α_1	6.7070	0.0323	6.6437	6.7703
β	0.1073	0.0387	0.0314	0.1832
$\log(\alpha_2)$	−0.5833	0.0271	−0.6364	−0.5302

Given are the parameter estimates, their standard error and the Wald confidence interval.

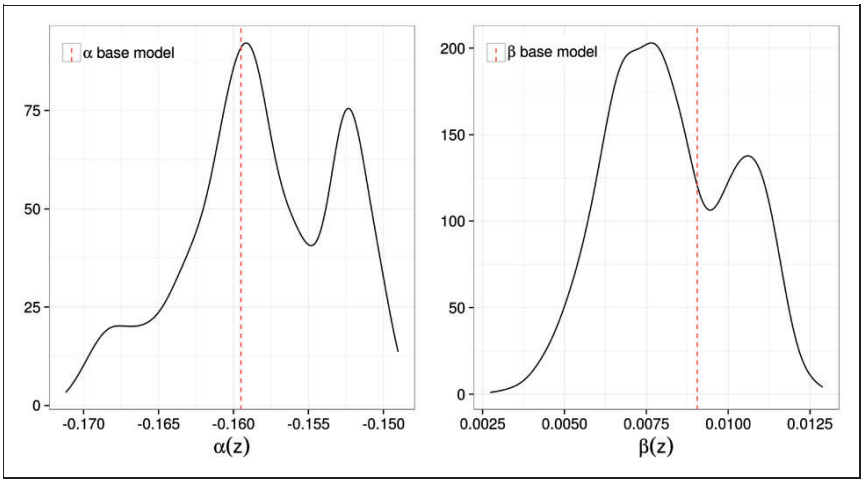


Figure 1. Kernel density estimates of the personalised parameter estimates for the ALSFRS.

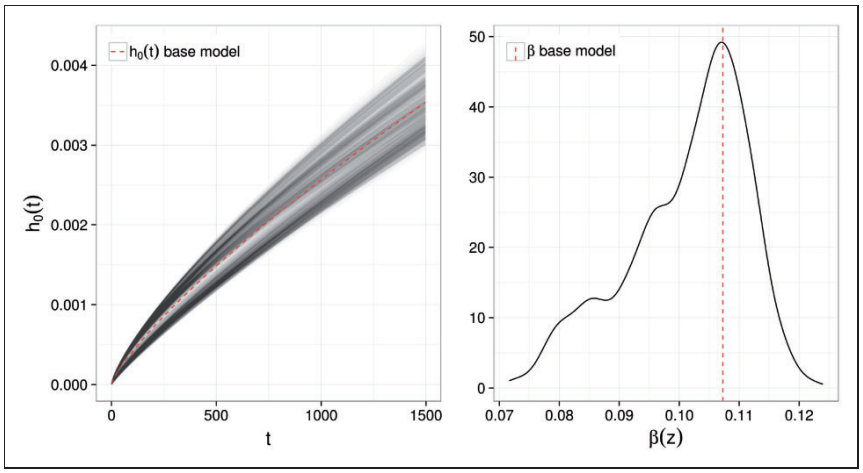


Figure 2. Distribution of the personalised parameter estimates for the survival time. The baseline hazard functions are given in the left panel; the kernel density estimate of the treatment effect estimate is given in the right panel.

baseline hazard, i.e. on α_1 and α_2. Instead of depicting the densities of the two baseline hazard parameters, in Figure 2, we show the baseline hazard curves. The baseline hazard varies for different patients, and there is a gap in the middle. The baseline hazard estimated from the base model lies close to that gap.

From the personalised models, we obtained the 'forest log-likelihoods' for both outcomes. For the Gaussian GLM with log-link and offset, the log-likelihood contribution for observation i is defined as

$$\begin{aligned} &l\Big((\text{ALSFRS}_6, \text{ALSFRS}_0, \mathbf{x})_i, \widehat{\vartheta}(\mathbf{z}_i)\Big) \\ &\quad = \Big(\text{ALSFRS}_{6i} - \exp\Big(\mathbf{x}_i^\top \widehat{\vartheta}(\mathbf{z}_i)\Big) \cdot \text{ALSFRS}_{0i}\Big)^2 \end{aligned} \tag{17}$$

with $\mathbf{x}_i = (1, x_{Ai})^\top$ and $\widehat{\vartheta}(\mathbf{z}_i) = (\hat{\alpha}(\mathbf{z}_i), \hat{\beta}(\mathbf{z}_i))^\top$. For the Weibull model, the log-likelihood contribution for observation i is

$$l\Big((\text{survival}, \mathbf{x})_i, \widehat{\vartheta}(\mathbf{z}_i)\Big) = \delta_i \log(\hat{\alpha}_2(\mathbf{z}_i)) - \delta_i \frac{\text{survival}_i - \mathbf{x}_i^\top \widehat{\vartheta}^*(\mathbf{z}_i)}{\hat{\alpha}_2(\mathbf{z}_i)} + \exp\left(\frac{\text{survival}_i - \mathbf{x}_i^\top \widehat{\vartheta}^*(\mathbf{z}_i)}{\hat{\alpha}_2(\mathbf{z}_i)}\right) \tag{18}$$

with $\mathbf{x}_i = (1, x_{iA})^\top$, $\widehat{\vartheta}^*(\mathbf{z}_i) = (\hat{\alpha}_1(\mathbf{z}_i), \hat{\beta}(\mathbf{z}_i))^\top$ and δ_i as the censoring indicator.

As can be seen in Figures 3 and 4, the forest log-likelihoods are higher than the log-likelihoods of the base models for both the ALSFRS and the survival time. The figures show the difference in log-likelihood between the forest and the corresponding base model. To show that this difference is not due to overfitting, we drew 50 samples from the base models, i.e. 50 parametric bootstrap samples for which the assumption holds that the intercept (or baseline hazard) and treatment effect are the same for all patients. ALSFRS values are drawn from a normal distribution truncated at 0 to assure positivity. (The effect of truncation is virtually negligible; only two observations had a truncation probability of more than 1%.) The survival times are drawn from a Weibull distribution censored at the originally observed censoring times (if exceeded). The differences in log-likelihoods for both ALSFRS and survival time are distributed close to 0, with a slight shift to the right, for the parametric bootstrap samples. The large difference in the ALS data supports the assumption that the base models are not ideal and personalised models are meaningful (the respective p-values are both 0). To approximately check the sub-hypotheses given in equations (11) and (12), we also computed log-likelihoods of the two forests that split only with respect to one of the partial score functions – either intercept (or baseline hazard) or treatment effect. For the ALSFRS, both the forest under H_1^α (computed with splitting only based on the partial score function with respect to the intercept α) and the forest under H_1^β (computed with splitting only based on partial score function with

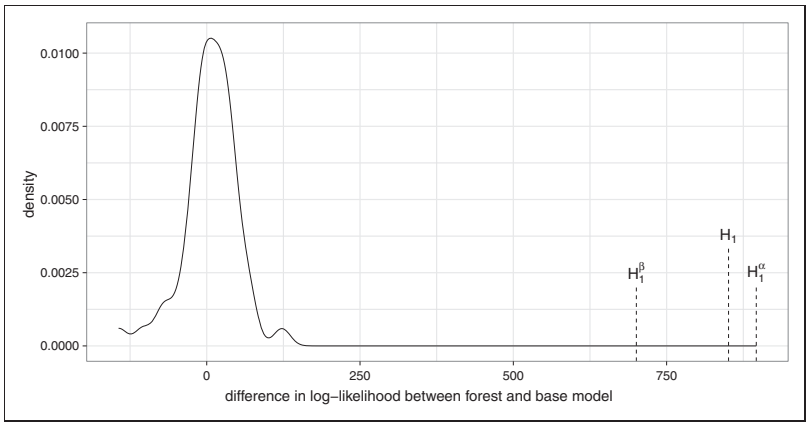


Figure 3. Difference in log-likelihoods between forest and base model using the original data (dashed lines; H_1, the usual forest; H_1^{α}, the forest that splits based on α; H_1^{β}, the forest that splits based on β) and using 50 samples simulated from the base model (density curve) for the ALSFRS outcome.

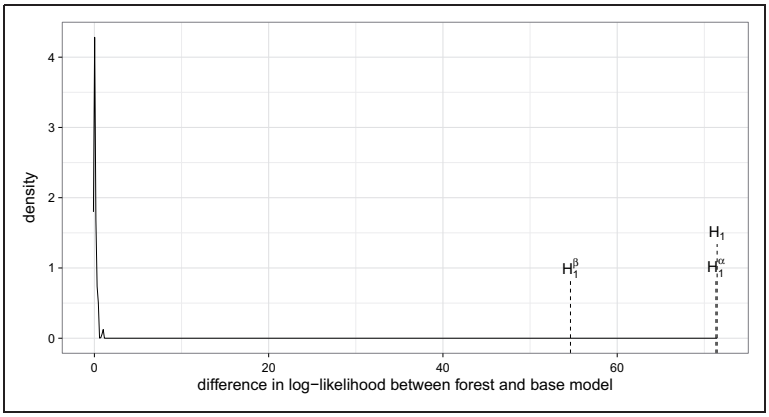


Figure 4. Difference in log-likelihoods between forest and base model using the original data (dashed lines; H_1, the usual forest; H_1^{α}, the forest that splits based on α; H_1^{β}, the forest that splits based on β) and using 50 samples simulated from the base model (density curve) for the survival outcome.

respect to the treatment effect β) lead to greatly improved models compared to the base model. The difference in log-likelihood between the forest under H_1^{α} and the base model is even greater than between the original forest (H_1) and the base model. For the survival time, the log-likelihoods of the original forest and the forest under H_1^{α} (based on splits in the partial score function with respect to the baseline hazard) are very close to each other. Splitting based only on the partial score function with respect to the treatment effect (H_1^{β}) already improves the log-likelihood but not as much as splitting based on both intercept and treatment effect (H_1). The good performances of the forests under H_1^{α} indicate that (a) there are no predictive patient characteristics, (b) all predictive patient characteristics are also prognostic or (c) the predictive nature of the predictive patient characteristics are so strong that it has enough impact on the structure of the partial score function with respect to α.

3.3 Dependence plots

The dependence plots as shown in Figures 5 and 6 can be obtained for any partitioning variable. Here we show the dependence plots for the four variables with the highest variable importance (see Section 3.4). For continuous variables, such as age, we show a scatter plot, as before. For categorical variables, such as the variable weakness, which indicates whether a patient suffers from muscle weakness (yes/no), boxplots giving the variation of $\beta(z)$ and a square representing $\bar{\beta}(z)$, i.e. the mean, are a meaningful way of representing the dependence between treatment effect and the given variable.

The most obvious pattern of the four graphs in Figure 5 is shown in Figure 5(d), in which the personalised treatment effects are plotted against the forced vital capacity (FVC). Patients with a low lung function (low FVC) are predicted to have a higher treatment effect than those with better lung function. The graph shows a relatively clear cut at approximately 3 L. This indicates that FVC is a predictive factor. For the time between disease onset and treatment start, the pattern is less clear. Patients with a short as well as those with a long time between disease onset and treatment start seem to benefit most. Also for the creatinine value, which indicates kidney function, only weak patterns are observed. The phosphorus balance is slightly negatively associated with the treatment effect.

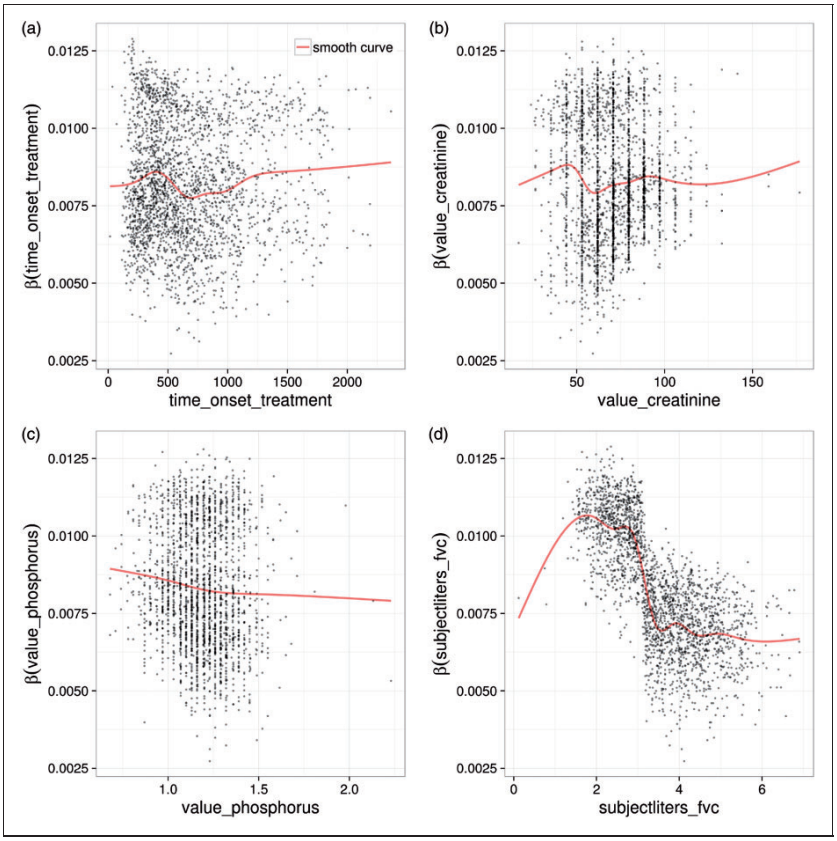


Figure 5. Dependence plots for the four patient characteristics with the highest variable importance from the ALSFRS forest. (a) Dependence plot for the time in days between disease onset and treatment start. (b) Dependence plot for the creatinine level in mmol/L. (c) Dependence plot for the phosphorus level in mmol/L. (d) Dependence plot for the forced vital capacity (volume of air in litres that can forcibly be blown out after full inspiration).

For the survival time, plotting only the treatment effect against a variable is not meaningful since the interpretation of the treatment effect depends on the shape of the baseline hazard. Therefore, we took a different approach in this case and show on the y-axis the difference in median survival between treatment and control intake. For example, a value of 70 means that based on the personalised model of this patient, the median survival is prolonged by 70 days if the patient takes Riluzole. The difference in median survival is denoted by $\Delta_{0.5}$. Any other quantile could be used as well since from the Weibull model, information on the entire estimated distribution in the two treatment groups is obtained. Taking the difference in medians makes sense because it is a measure on the scale of the outcome, just as the treatment effect in a linear model, which is the difference in means. The shape of $\Delta_{0.5}$ when plotted against age shows a strong pattern that indicates that age is a predictive factor (see Figure 6). The treatment efficacy increases with age until about 55 years and then flattens. The difference in median survival slightly increases with the days between disease onset and start of treatment in the beginning, but decreases again after about 1000 days. Patients who suffer from weakness have a greater variance in their benefit from Riluzole. Tall patients are predicted to benefit little on average.

3.4 Variable importance

Figures 7 and 8 show the variable importance of each split variable. Figure 7 suggests that the time between disease onset and start of treatment plays the most important role for the personalised models. The time between disease

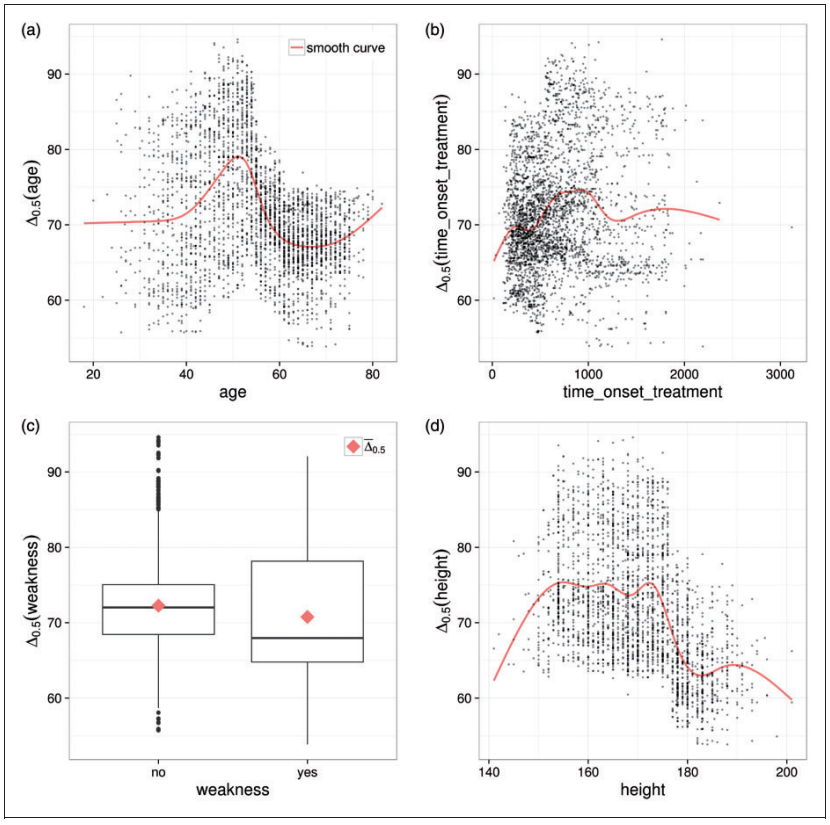


Figure 6. Dependence plots for the four patient characteristics with the highest variable importance from the survival time forest. (a) Dependence plot for the age. (b) Dependence plot for the time in days between disease onset and treatment start. Outlier has been omitted in the estimation of the smooth curve. (c) Dependence plot for the weakness indicator. (d) Dependence plot for the height.

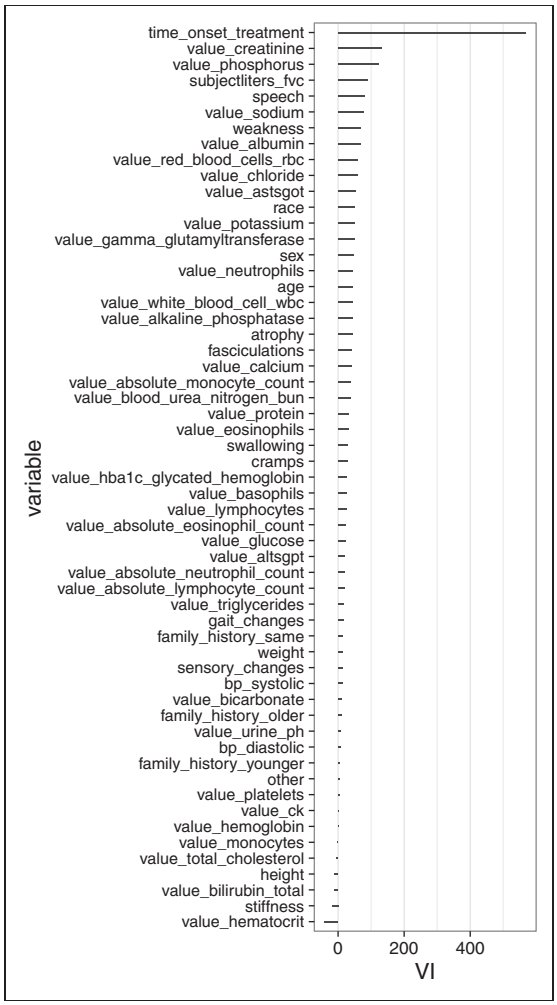


Figure 7. Variable importances of all split variables used for the ALSFRS forest.

onset and start of treatment, the FVC, and the phosphorus balance have been shown to be the most important variables for stratified models,[5] which is underlined by this analysis. The time between disease onset and start of treatment contains information on the state of disease progression for patients in the trial. If the disease onset and the start of treatment are far apart, the patient is likely to have a slow progression.[28] Also Riluzole has been shown to not be effective when the disease is already far progressed.[1] Thus, it is not surprising that this variable is selected as an important variable.

For the Riluzole effect on the survival time, the patient's age and again the time between onset and treatment start play a role. Both variables have been identified before[5] as important factors for survival time.

4 Discussion

Model-based forests can find important predictive and prognostic patient characteristics and – more importantly – via the personalised models provide the possibility to predict the counterfactual individual treatment

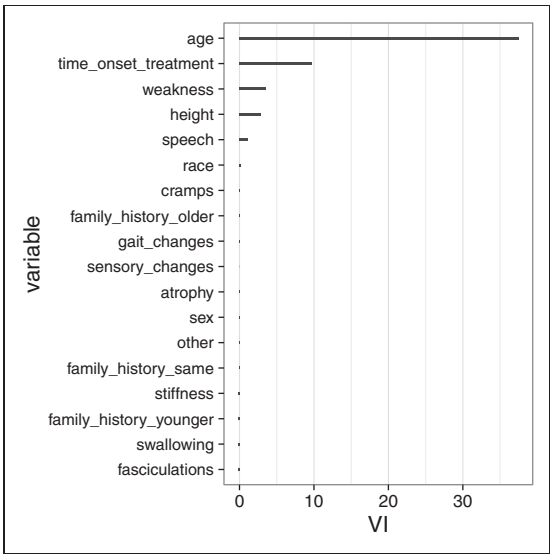


Figure 8. Variable importances of all split variables used for the survival time forest.

effect of a future patient. The personalised models allow a shift from standardised medicine back to personalised medicine, but this time in a controlled way by using statistical principles. Through analysis of the PRO-ACT data and simulations (see Appendix 3), we showed that personalised models can perform better than the standard global model if there are differences in treatment effect between patients. If there is no difference, the performance of the methods is about the same. In our performance checks, we focused on the fit of the model to the data based on the log-likelihood. Performance of the method for new patients was studied using simulations.

The proposed method is applicable to clinical trial data where treatment is randomised. In our analysis of the PRO-ACT data, we included several clinical trials for which we have no knowledge about inclusion criteria or any other details of the study protocol as this information is not given out in order to anonymise data. This could possibly lead to confounding issues. As there is interest in methodology for when treatment is not randomised, we included a small simulation study on this topic in Appendix 4. The results seem promising in the case where the patient characteristic that impacts the treatment assignment is not the predictive factor. However, there is a bias when a patient characteristic is predictive and also impacts treatment assignment. Further work in the area of observational trials is needed where, e.g. adjustment methods such as propensity scoring[29] could be of use.

The presented methods are based on tree-based subgroup analyses but go a step further. Not only are subgroups identified and the treatment effect within each group estimated, but many slightly varying trees are used to retrieve a measure of similarity between patients. On this basis, a model is computed in which more similar patients are weighted higher. The personalised models provide point estimates for the treatment effect. When the individual treatment effects are plotted against patient characteristics, researchers can determine whether the patient characteristics are predictive factors and in what way the patient characteristics and the treatment interact. For ALS patients, the FVC value was predictive for the ALSFRS, and the patient's age and height were predictive for survival. The next step would be to generate hypotheses from these findings and plan a study to test these. Our method offers a promising means of providing individual treatment effect predictions and can be applied to any clinical trial data where baseline patient characteristics are available.

All results were obtained solely using open-source implementation software (see Section 5), which provides easy access to the methods.

5 Computational details

The code for data preprocessing of the PRO-ACT data is available in the TH.data package.[30] The source code for the full analyses is available on https://github.com/HeidiSeibold/personalised_medicine. Implementation of all methods discussed in this article is based on the R partykit package (version 1.0-2).[31] Other R packages used were sandwich (2.3-3),[32,33] survival (2.38-1),[34] eha (2.4-2)[35] and ggplot2 (2.0.0).[36] All computations were conducted in the R system for statistical computing (version 3.2.0).[37]

Acknowledgements

We thank Karen A. Brune for improving the language.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Heidi Seibold and Torsten Hothorn were financially supported by the Swiss National Science Foundation (Grant 205321_163456).

References

1. European Medicines Agency. Riluzole Zentiva: EPAR summary for the public, http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Summary_for_the_public/human/002622/WC500127609.pdf (2012, accessed 28 January 2017).
2. Atassi N, Berry J, Shui A, et al. The PRO-ACT database: Design, initial analyses, and predictive features. *Neurology* 2014; **83**: 1719–1725.
3. Weisberg HI. What next for randomised clinical trials? *Significance* 2015; **12**: 22–27.
4. Italiano A. Prognostic or predictive? It's time to get back to definitions! *J Clin Oncol* 2011; **29**: 4718–4718.
5. Seibold H, Zeileis A and Hothorn T. Model-based recursive partitioning for subgroup analyses. *Int J Biostat* 2016; **12**: 45–63.
6. Ciampi A, Negassa A and Lou Z. Tree-structured prediction for censored survival-data and the Cox model. *J Clin Epidemiol* 1995; **48**: 675–689.
7. Kehl V and Ulm K. Responder identification in clinical trials with censored data. *Comput Stat Data Anal* 2006; **50**: 1338–1355.
8. Dusseldorp E and Van Mechelen I. Qualitative interaction trees: A tool to identify qualitative treatment-subgroup interactions. *Stat Med* 2013; **33**: 219–237.
9. Loh WY, He X and Man M. A regression tree approach to identifying subgroups with differential treatment effects. *Stat Med* 2015; **34**: 1818–1833.
10. Tian L, Alizadeh AA, Gentles AJ, et al. A simple method for estimating interactions between a treatment and a large number of covariates. *J Am Stat Assoc* 2014; **109**: 1517–1532.
11. Foster JC, Taylor JMG, Kaciroti N, et al. Simple subgroup approximations to optimal treatment regimes from randomized clinical trial data. *Biostatistics* 2015; **16**: 368–382.
12. Zhang B, Tsiatis AA, Davidian M, et al. Estimating optimal treatment regimes from a classification perspective. *Stat* 2012; **1**: 103–114.
13. Breiman L. Random forests. *Mach Learn* 2001; **45**: 5–32.
14. Zeileis A, Hothorn T and Hornik K. Model-based recursive partitioning. *J Comput Graph Stat* 2008; **17**: 492–514.
15. Tutz G. *Regression for categorical data*. New York: Cambridge University Press, 2012.
16. Loh WY. Regression trees with unbiased variable selection and interaction detection. *Stat Sin* 2002; **12**: 361–386.
17. Hothorn T, Hornik K and Zeileis A. *ctree: Conditional inference trees*, Vignette R package partykit version 1.1-1, https://CRAN.R-project.org/web/packages/partykit/vignettes/ctree.pdf (2016, accessed 28 January 2017).
18. Hothorn T, Hornik K, Van de Wiel MA, et al. A Lego system for conditional inference. *Am Stat* 2006; **60**: 257–263.
19. Hothorn T, Hornik K and Zeileis A. Unbiased recursive partitioning: A conditional inference framework. *J Comput Graph Stat* 2006; **15**: 651–674.
20. Zeileis A and Hornik K. Generalized M-fluctuation tests for parameter instability. *Stat Neerland* 2007; **61**: 488–508.

21. Hothorn T, Lausen B, Benner A, et al. Bagging survival trees. *Stat Med* 2004; **23**: 77–91.
22. Meinshausen N. Quantile regression forests. *J Mach Learn Res* 2006; **7**: 983–999.
23. Lin Y and Jeon Y. Random forests and adaptive nearest neighbors. *J Am Stat Assoc* 2006; **101**: 578–590.
24. Hothorn T and Zeileis A. Transformation forests. Technical Report, arXiv 1701.02110, v1, https://arxiv.org/abs/1701.02110 (2017, accessed 28 January 2017).
25. Hastie T, Tibshirani R and Friedman J. *The elements of statistical learning*. 2nd ed. Berlin: Springer-Verlag, 2009.
26. Strobl C, Boulesteix AL, Kneib T, et al. Conditional variable importance for random forests. *BMC Bioinform* 2008; **9**: 307.
27. Brooks BR, Sanjak M, Ringel S, et al. The amyotrophic lateral sclerosis functional rating scale – assessment of activities of daily living in patients with amyotrophic lateral sclerosis. *Archiv Neurol* 1996; **53**: 141–147.
28. Hothorn T and Jung HH. RandomForest4life: A random forest for predicting ALS disease progression. *Amyotrop Lateral Scler Frontotemp Degen* 2014; **15**: 444–452.
29. Rosenbaum PR and Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**: 41–55.
30. Hothorn T. *TH.data: TH's data archive*, R package version 1.0-7, https://CRAN.R-project.org/package=TH.data (2016, accessed 28 January 2017).
31. Hothorn T and Zeileis A. partykit: A modular toolkit for recursive partytioning in R. *J Mach Learn Res* 2015; **16**: 3905–3909.
32. Zeileis A. Econometric computing with HC and HAC covariance matrix estimators. *J Stat Softw* 2004; **11**: 1–17.
33. Zeileis A. Object-oriented computation of sandwich estimators. *J Stat Softw* 2006; **16**: 1–16.
34. Therneau TM. *A package for survival analysis in S*, Version 2.40-1, https://CRAN.R-project.org/package=survival (2016, accessed 28 January 2017).
35. Broström G. *eha: Event history analysis*, R package version 2.4-4, https://CRAN.R-project.org/package=eha (2016, accessed 28 January 2017).
36. Wickham H. *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag, 2009.
37. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2016.
38. Strasser H and Weber C. On the asymptotic theory of permutation statistics. *Mathem Method Stat* 1999; **8**: 220–250.
39. Strobl C, Boulesteix AL, Zeileis A, et al. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinform* 2007; **8**: 25.
40. Foster JC, Taylor JM and Ruberg SJ. Subgroup identification from randomized clinical trial data. *Stat Med* 2011; **30**: 2867–2880.

Appendix I

Notation

$b = 1, \ldots, B$	subgroup number
$\mathcal{B}$	subgroup
$i, k = 1, \ldots, N$	observation number
$j = 1, \ldots, J$	patient characteristics number
ℓ	log-likelihood
$\mathcal{L}$	data sample ($\mathcal{L}_t$ training sample tree t, $\mathcal{L}_t^c$ out-of-bag sample tree t)
$\mathcal{M}$	model
S	score (derivative of the log-likelihood $s = \partial l / \partial \vartheta$, partial derivatives are denoted by s_α, s_β, etc.)
$t = 1, \ldots, T$	tree number
w	weight
$\mathbf{X}$	model covariates (including treatment indicator X_A)
Y	response
$\mathbf{Z}$	patient characteristics
α	intercept
β	treatment effect
ϑ	model parameters $\vartheta = (\alpha, \beta, \ldots)^\top$

Appendix 2

Split algorithm in detail

In the following, the split algorithm in model-based recursive partitioning is explained. The split procedure starts with all N data points. In nodes other than the root node, the size of the data set depends on the previous splits. For notational simplicity, we describe the split procedure in the root node, i.e. for patients $i = 1, \ldots, N$.

- Compute prespecified (parametric) model $\mathcal{M}((Y, \mathbf{X}), \vartheta)$. Estimate $\widehat{\vartheta}$ by maximising the log-likelihood

$$\widehat{\vartheta} = \underset{\vartheta}{\operatorname{argmax}}\, \ell\left((Y, \mathbf{X}), \vartheta\right)$$

  or equivalently by solving

$$\sum_{i=1}^{N} s((y, \mathbf{x})_i, \vartheta) = 0$$

  for ϑ.
- Compute associated empirical estimating function (residuals)

$$\boldsymbol{s} = \begin{pmatrix} s_{\hat{\alpha}}((y, \mathbf{x})_1, \widehat{\vartheta}) & s_{\hat{\beta}}((y, \mathbf{x})_1, \widehat{\vartheta}) \\ s_{\hat{\alpha}}((y, \mathbf{x})_2, \widehat{\vartheta}) & s_{\hat{\beta}}((y, \mathbf{x})_2, \widehat{\vartheta}) \\ \vdots & \vdots \\ s_{\hat{\alpha}}((y, \mathbf{x})_N, \widehat{\vartheta}) & s_{\hat{\beta}}((y, \mathbf{x})_N, \widehat{\vartheta}) \end{pmatrix}$$

- Perform tests of independence between residuals (partial score vectors) $\boldsymbol{s}_{\hat{\alpha}}$ as well as $\boldsymbol{s}_{\hat{\beta}}$ and partitioning variables Z_j.

$$\begin{aligned} H_0^{\alpha,j}: &\quad \boldsymbol{s}_{\hat{\alpha}}((Y, \mathbf{X}), \widehat{\vartheta}) \quad \perp \quad Z_j \\ H_0^{\beta,j}: &\quad \boldsymbol{s}_{\hat{\beta}}((Y, \mathbf{X}), \widehat{\vartheta}) \quad \perp \quad Z_j \quad j = 1, \ldots, J \end{aligned}$$

For the tests, we use permutation testing with the linear statistic

$$T_j = \sum_{i \in \mathcal{B}_b} g_j(Z_{ji}) \cdot \left(\mathbf{s}_{\alpha}((y, \mathbf{x})_i, \widehat{\vartheta}),\ \mathbf{s}_{\beta}((y, \mathbf{x})_i, \widehat{\vartheta})\right)$$

The transformation function g depends on the scale of the variable Z_j. If Z_j is numeric, then $g_j(z_{ji}) = z_{ji}$. If Z_j is categorical with K categories, then $g_j(z_{ji}) = e_K(z_{ji}) = (I(z_{ji} = 1), \ldots, I(z_{ji} = K))^{\top}$, i.e. g_j is the unit vector of length K, where the element that corresponds to the value of z_{ji} is one. Note that T_j is two-dimensional for numeric patient characteristics and $2 \times K$-dimensional for categorical patient characteristics. If there are missing values in Z_j, the observations are excluded from the sum so that we actually sum over all observations $i \in \mathcal{B}_b$, except for the observations in $\mathcal{B}_b$, where Z_j is missing. The standardised test statistic is the Pearson correlation coefficient

$$c(t_j, \mu_j, \Sigma_j) = \left| \frac{(t_j - \mu_j)}{\sqrt{(\Sigma_j)}} \right|$$

if Z_j is numeric and otherwise

$$c(\mathbf{t}_j, \mu_j, \Sigma_j) = \max_{k=1,\ldots,K} \left| \frac{(\mathbf{t}_j - \mu_j)_k}{\sqrt{(\Sigma_j)_{kk}}} \right|$$

The conditional expectation μ_j and covariance Σ_j can be derived as in Strasser and Weber.[38] The smallest p-value corresponds to the largest discrepancy from the model assumption that intercept and treatment parameter are the same for all patients in the given node/subgroup.

- If any Bonferroni-adjusted p-value is lower than the significance level, select the partitioning variable Z_{j*} that has the highest association (lowest p-value) to any of the residuals relevant for the split.
- Select as split point the point that results in the largest discrepancy between score functions in the two resulting subgroups. The discrepancy can be measured by the linear statistic

$$T_{j*}^{k} = \sum_{i \in \mathcal{B}_{1k}} \mathbf{s}_i$$

where $\mathcal{B}_{1k}$ here is the first of the two new subgroups that are defined by splitting in split point k of variable Z_{j*}. The split point is then chosen as follows:

$$k* = \operatorname*{argmin}_{k} c(t_{j*}^{k}, \mu_{j*}^{k}, \Sigma_{j*}^{k})$$

Appendix 3
Empirical evaluation

To check whether the proposed method can recover smooth treatment effect functions, we evaluated its performance on artificial data. To do so, we simulated data from a normal linear regression model. We simulated 10 correlated patient characteristics, where only one is in a non-linear interaction with the treatment. In the following, we compare the log-likelihood of our method to the log-likelihood of the true underlying model and the naive model that assumes an overall applicable treatment effect (Appendix 3.1) and show the predicted treatment effects in dependence plots (Appendix 3.2) and the variable importances of the true predictive factor and the noise variables (Appendix 3.3).

We simulated 600 patients, half of which were treated ($x_A = 1$) and half of which were untreated ($x_A = 0$). The 10 partitioning variables $\mathbf{Z}$ are normally distributed

$$\mathbf{Z} \sim \mathcal{N}_{10}(\mathbf{0}, \Sigma_Z) \tag{19}$$

and correlated with the covariance matrix

$$\sum_Z = \begin{pmatrix} 1 & 0.2 & \cdots & 0.2 \\ 0.2 & 1 & \cdots & 0.2 \\ \vdots & \vdots & \ddots & \vdots \\ 0.2 & 0.2 & \cdots & 1 \end{pmatrix}$$

The primary outcome depends on treatment and partitioning variables as follows:

$$Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z} \sim \mathcal{N}(1.9 + 0.2 \cdot x_A + 3 \cdot \cos(z_1) \cdot x_A, 1) \tag{20}$$

In this example, the true model parameters are defined as follows:

$$\begin{aligned} \alpha(\mathbf{z}) &= 1.9 \\ \beta(\mathbf{z}) &= 0.2 + 0.3 \cdot \cos(z_1) \end{aligned} \tag{21}$$

This means that the treatment effect depends on the value of z_1 and this dependency has the form of a cosine function (see Figure 9).

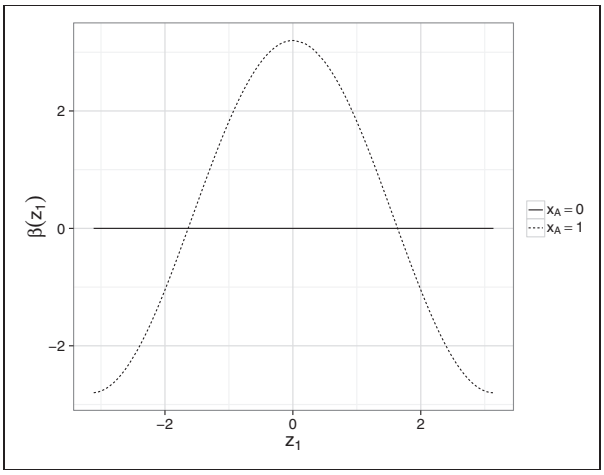


Figure 9. True treatment effect in given simulated data.

3.1 Comparison of models by comparing the log-likelihood

To compare our method with (a) a correctly specified model taking into account the main effects of x_A and $\cos(z_1)$ as well as the interaction of x_A and $\cos(z_1)$ and (b) a simple linear model including only the treatment x_A as a covariate, we drew 100 learning samples and 100 test samples using the data simulation procedure explained above and computed the out-of-sample log-likelihoods (i.e. based on the test data) for the models after applying them to each of the 100 learning data sets. The log-likelihood contributions

$$\ell\Big((y,\mathbf{x})_i,\widehat{\vartheta}(\mathbf{z}_i)\Big)=\Big(y_i-\mathbf{x}_i^\top\widehat{\vartheta}(\mathbf{z}_i)\Big)^2 \tag{22}$$

with $\mathbf{x}_i=(1,x_{iA})^\top$ and $\widehat{\vartheta}(\mathbf{z}_i)=(\hat{\alpha}(\mathbf{z}_i),\hat{\beta}(\mathbf{z}_i))^\top$ are taken from the personalised models of our method (see Section 2.5). Note that for the simple linear model the log-likelihood contributions are defined as above, but only with constant parameters, for the fully specified model $\mathbf{x}_i=(1,x_A,\cos(z_1),x_A\cdot\cos(z_1))^\top$ and $\hat{\vartheta}=(\hat{\alpha},\hat{\beta}_A,\hat{\beta}_{\cos(z_1)},\hat{\beta}_{A,\cos(z_1)})^\top$.

The log-likelihoods of our method are higher than the log-likelihoods of the simple and incorrect linear model and lower than the log-likelihoods of the correctly specified model (Figure 10). Therefore, we conclude that our method performs reasonably well.

3.2 Dependence plots

For the same 100 simulated test data sets as above, we obtained the dependence plots. Figure 11 shows two dependence plots in which all 100 simulations are combined by layering them on top of each other. The dependence plot of partitioning variable z_1 (Figure 11(a)) shows a curve that is fairly similar to that of Figure 9, except that the effect is shrunken towards 0. Note that with a larger sample or differently tuned parameters (e.g. larger trees), one could get better results for the extreme treatment effects. As expected, for partitioning variables z_2 to z_{10}, there is only random fluctuation around 0 (see as an example Figure 11(b), which shows the dependence plot for z_2).

3.3 Variable importance

Variable importances for one simulated data set are shown in Figure 12. As expected, partitioning variable z_1 is the only variable with a clearly positive variable importance. Even though all partitioning variables are correlated, the method was able to distinguish between the correlation and predictive effect.

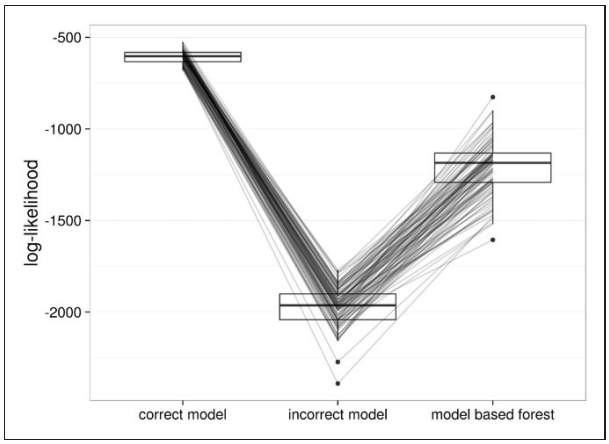


Figure 10. Out-of-sample log-likelihoods obtained from the three models. Each line represents one simulated data set.

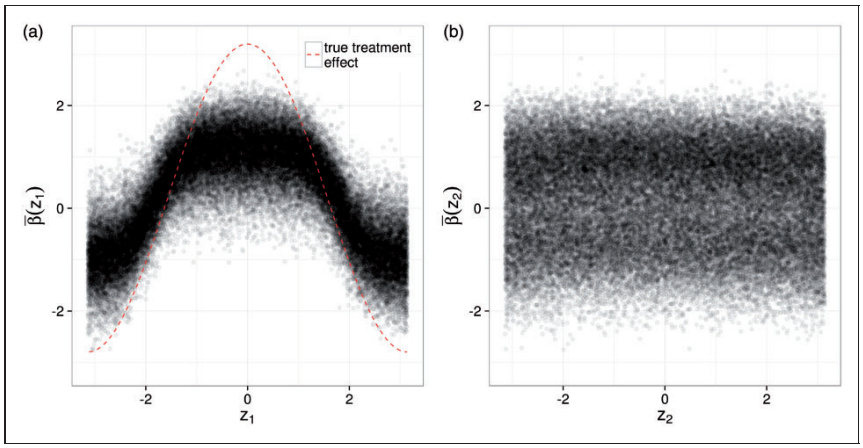


Figure 11. Joint dependence plots of all 100 simulations. (a) Dependence plot for z1. (b) Dependence plot for z2.

3.4 Comparison to regular forest

Using the same simulated data sets, we compared the personalised models to the output of a conditional inference forest, a random forest of conditional inference trees.[19,39] For the random forest, we computed the treatment effect for each patient by:

$$\hat{\beta}(\mathbf{z}_i) = \hat{y}_i^{(1)} - \hat{y}_i^{(0)} \tag{23}$$

where $\hat{y}_i^{(1)}$ is the predicted outcome when patient i is treated, i.e. $x_{Ai}=1$, and $\hat{y}_i^{(0)}$ is the predicted outcome when patient i is not treated, i.e. $x_{Ai}=0$. This is a common strategy and used, for example, in the first step of the Virtual Twins algorithm.[40] Note that equation (23) can only be interpreted as personalised treatment effect for continuous

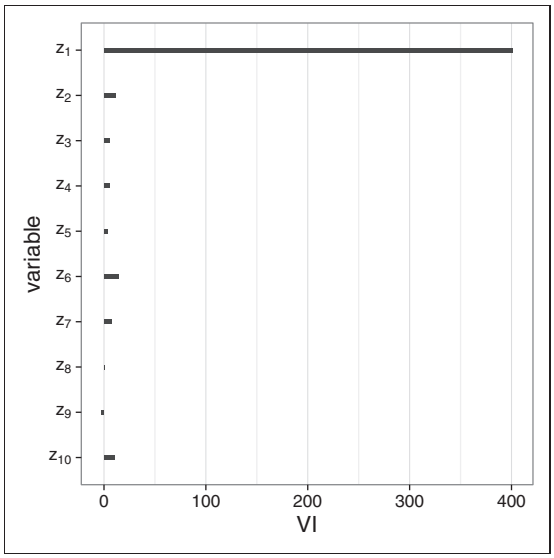


Figure 12. Variable importances of the predictive factor z_1 and noise variables z_2 to z_{10}.

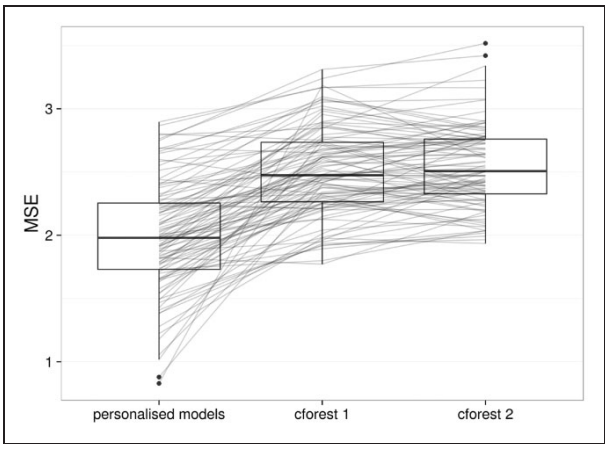


Figure 13. Mean-squared error between true treatment effect and estimates. Comparison of personalised models and conditional inference forest (without and with treatment × patient characteristics interactions). (a) Dependence plot for z_1 with $P(A) = f(z_1)$. (b) Dependence plot for z_2 with $P(A) = f(z_1)$. (c) Dependence plot for z_1 with $P(A) = 1 - f(z_1)$. (d) Dependence plot for z_2 with $P(A) = 1 - f(z_1)$. (e) Dependence plot for z_1 with $P(A) = f(z_2)$. (f) Dependence plot for z_2 with $P(A) = f(z_2)$. (g) Dependence plot for z_1 with $P(A) = 1 - f(z_2)$. (h) Dependence plot for z_2 with $P(A) = 1 - f(z_2)$.

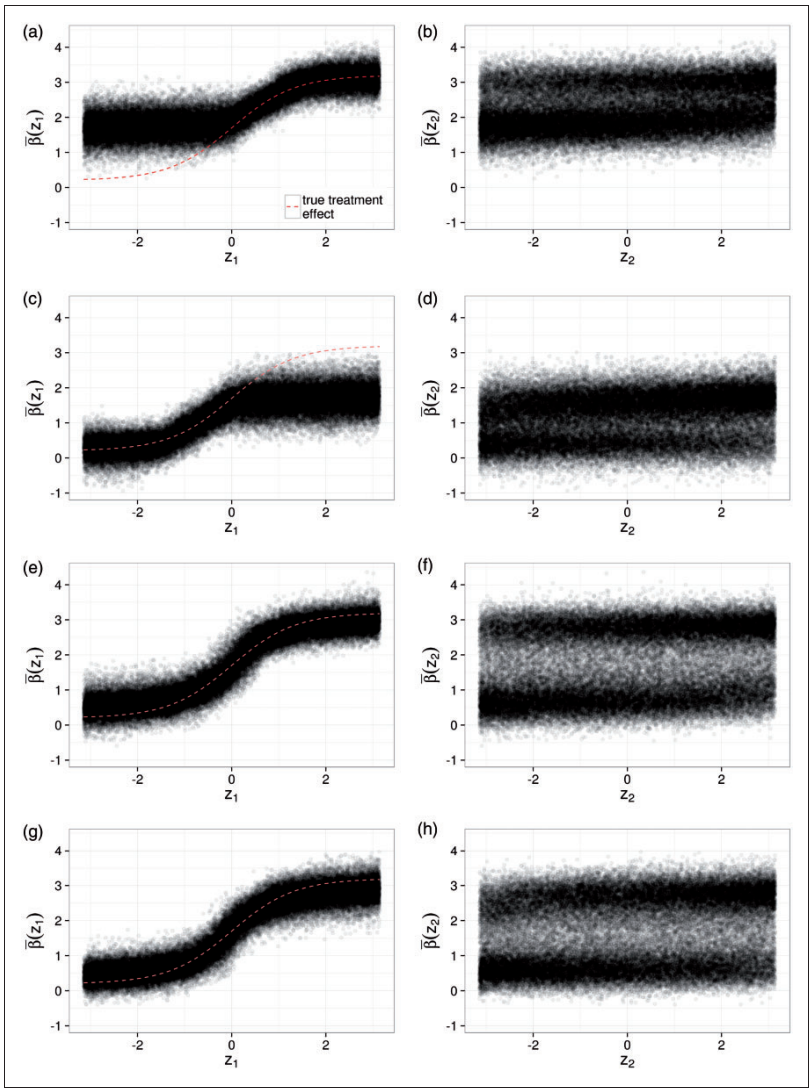


Figure 14. Joint dependence plots of all 100 simulations. (a) Dependence plot for z1 with P(A) = f(z1). (b) Dependence plot for z2 with P(A) = f(z1). (c) Dependence plot for z1 with P(A) = 1 – f(z1). (d) Dependence plot for z2 with P(A) = 1 – f(z1). (e) Dependence plot for z1 with P(A) = f(z2). (f) Dependence plot for z2 with P(A) = f(z2). (g) Dependence plot for z1 with P(A) = 1 – f(z2). (h) Dependence plot for z2 with P(A) = 1 – f(z2).

outcomes. In all other situations, a model-based approach allowing a clear treatment effect parameter to be included is mandatory.

We looked at two versions of the random forest: One, where we include all patient characteristics and the treatment indicator as split variables (cforest 1) and one, where we additionally include the interaction terms between each patient characteristic and the treatment indicator (cforest 2), i.e. $x_A \cdot z_j$ $(j = 1, \ldots, 10)$.

The distribution of mean-squared error (MSE) between the true treatment effect and the estimates for the 100 simulated data sets is shown in Figure 13. The personalised models outperform the random forests on almost all data sets generated. The inclusion of the interaction as split variable does not improve the random forest, which can be expected as random forests are good at finding interactions by design. The average MSE over 100 simulations is 1.99 for the personalised models, 2.51 for the first version of the random forest and 2.54 for the second version.

Appendix 4

Impact of non-randomised treatment

Even though the personalised models were developed for the application in randomised clinical trials, we acknowledge that randomisation is not always possible and the interest for methods that can be used in observational studies is high. Here, we investigated the behaviour of the method when the treatment assignment depends on a patient characteristic. We simulated data similar to Appendix 3, but instead of a cosine function we now use the logistic function

$$Y|\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z} \sim \mathcal{N}(1.9+0.2\cdot x_A+3\cdot f(z_1)\cdot x_A,1) \tag{24}$$

with

$$f(z)=\frac{1}{1+\exp(-1.5\cdot z)} \tag{25}$$

and the treatment assignment is not randomised but depends on a patient characteristic. We consider four scenarios for the probability of receiving treatment A:

$$P(x_A=1|\mathbf{Z}=\mathbf{z}) = \begin{cases} f(z_1) & \text{scenario 1} \\ 1-f(z_1) & \text{scenario 2} \\ f(z_2) & \text{scenario 3} \\ 1-f(z_2) & \text{scenario 4} \end{cases} \tag{26}$$

In scenarios 1 and 2, the prognostic factor z_1 influences the probability of being treated with treatment A. In scenarios 2 and 3 z_2, a patient characteristic with no influence on the outcome, influences the probability of being treated with treatment A. Dependence plots are shown in Figure 14. Scenarios 1 and 2 lead to a bias in the estimation of the treatment effect (Figure 14(a) to (d)), but for scenarios 3 and 4, there does not seem to be a problem. Therefore, in the cases where the patient characteristic that influences the treatment assignment is not a predictive factor, this simple simulation study reveals no problems.

PAPER III

Generalised Linear Model Trees with Global Additive Effects

Heidi Seibold, Torsten Hothorn, Achim Zeileis

Accepted in *Advances in Data Analysis and Classification*, 2018.

Generalised Linear Model Trees with Global Additive Effects

Heidi Seibold
University of Zurich
LMU Munich

Torsten Hothorn
University of Zurich

Achim Zeileis
Universität Innsbruck

Abstract

Model-based trees are used to find subgroups in data which differ with respect to model parameters. In some applications it is natural to keep some parameters fixed globally for all observations while asking if and how other parameters vary across subgroups. Existing implementations of model-based trees can only deal with the scenario where all parameters depend on the subgroups. We propose partially additive linear model trees (PALM trees) as an extension of (generalised) linear model trees (LM and GLM trees, respectively), in which the model parameters are specified a priori to be estimated either globally from all observations or locally from the observations within the subgroups determined by the tree. Simulations show that the method has high power for detecting subgroups in the presence of global effects and reliably recovers the true parameters. Furthermore, treatment-subgroup differences are detected in an empirical application of the method to data from a mathematics exam: the PALM tree is able to detect a small subgroup of students that had a disadvantage in an exam with two versions while adjusting for overall ability effects.

Keywords: subgroup analysis, model-based recursive partitioning, GLM, tree.

1. Introduction

Model-based recursive partitioning (Zeileis, Hothorn, and Hornik 2008) is used to partition data into groups that differ in terms of the parameters in the model. The method can be applied, for example, to find subgroups in a clinical trial which differ in terms of treatment effect on a health score (e.g. Seibold, Zeileis, and Hothorn 2016) or areas in a city which differ in terms of the influence of square metres on the rent price. Sometimes there are parameters in the model that one wants to fix for all groups, e.g. the effect of smoking on the health outcome in the clinical trial or the effect of inflation/deflation on rent prices. This, however, is not possible in model-based recursive partitioning as described in Zeileis et al (2008). Here we propose an algorithm called PALM tree that is similar to model-based recursive partitioning but allows fixing parameters over all groups, i.e. only some parameters depend on the tree structure.

There have been several developments in the past years toward the direction of combining models and trees, where one part of the model follows a tree structure and one part does not. The Simultaneous Threshold Interaction Modeling Algorithm (STIMA, Dusseldorp, Conversano, and Van Os 2010) starts off with a main effects model and adds interactions

based on a tree. Fokkema, Smits, Zeileis, Hothorn, and Kelderman (2017) proposed GLMM tree, a method that is similar to PALM tree, but is used to fix random effects in a generalised linear mixed-effects model (GLMM) instead of – as in PALM tree – further fixed effects. Other approaches going in the direction of GLMM tree are RE-EM tree (Sela and Simonoff 2012) and MERT (Hajjem, Bellavance, and Larocque 2011).

In the literature on subgroup analyses for the estimation of treatment effects, special tree-based procedures have been proposed (see, e.g. Doove, Dusseldorp, Van Deun, and Van Mechelen 2014). These methods are commonly used in the analysis of clinical trials, but are equally relevant in contexts such as marketing studies evaluating different marketing strategies or studies on website user behaviour, where users are randomly served one of two website versions (A/B testing). Sies and Van Mechelen (2017) review some of the methods in a setting where there are some model covariates with fixed parameters across all subgroups and varying treamtent effect. One promising method in this review is a method by Zhang, Tsiatis, Davidian, Zhang, and Laber (2012) which estimates rules of optimal treatment for each patient subgroup (optimal treatment regimes).

The following sections unfold as follows: In Section 2 we will first describe GLMs and GLM trees as the basics needed for PALM trees and then go into how PALM trees are computed. Furthermore we will show how model-based trees (LM trees, GLM trees and PALM trees) can be used for finding subgroups with differential treatment effects. In Section 3 we will show the results of a simulation study in which we compare LM tree, PALM tree, STIMA and the optimal treatment regime method by Zhang et al (2012). In Section 4 we will apply the PALM tree to data of a mathematics exam, where the endpoint is performance in the exam, the "treatment" is the student group (early morning or late group) and the known prognostic factor is the performance in online tests the students participate in during the semester. Finally we will discuss strengths and limitations of model-based trees in general and PALM trees in particular.

2. Methods

In this section we first describe the basics needed for PALM trees – GLMs and GLM trees – and then introduce PALM trees and how GLMs and GLM trees are used in the PALM tree algorithm. We focus on GLMs since LMs are a special case of GLMs.

2.1. Basics: GLMs and GLM trees

GLMs

GLMs model the expected response $\mu = \mathbb{E}(y)$ given the covariates $\mathbf{x}$. To fix notation we write the GLM as $g(\mu) = \mathbf{x}^\top \boldsymbol{\beta}$ where g denotes the link function and $\mathbf{x}^\top \boldsymbol{\beta}$ the linear predictor with coefficient vector $\boldsymbol{\beta}$. The coefficients are estimated by maximising the log-likelihood. The observation-wise log-likelihood contributions are denoted by $l((y, \mathbf{x})_i, \boldsymbol{\beta})$ with $i = 1, \ldots, n$ indexing the i-th observation and l is defined depending on the appropriate exponential family chosen for the GLM (Gaussian, Poisson, etc.).

In the following we will make use of two refinements commonly used in GLMs: (a) interactions and (b) offsets. Interactions are effects combinding two or more covariates and can be

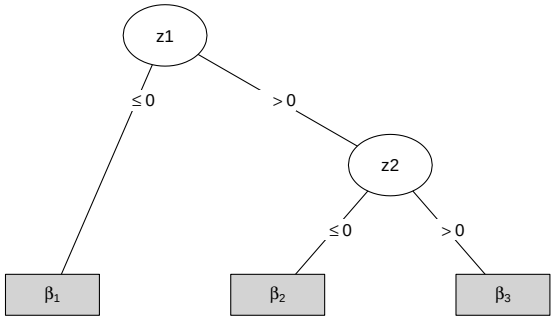


Figure 1: Example of a model-based tree.

employed to establish subgroup-specific coefficient vectors in a single model:

$$g(\mu) = \tilde{\mathbf{x}}^\top \tilde{\boldsymbol{\beta}} = I(\text{subgroup}_1) \cdot \mathbf{x}^\top \boldsymbol{\beta}_1 \; + I(\text{subgroup}_2) \cdot \mathbf{x}^\top \boldsymbol{\beta}_2 \; + \ldots \tag{1}$$

where $I(\text{subgroup}_j)$ equals 1 for observations in the j-th subgroup and 0 for others. The combined coefficient vector is simply $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \ldots)^\top$

Offsets in GLMs are useful for incorporating additional terms whose effects are known or fixed into the linear predictor :

$$g(\mu) = \mathbf{x}^\top \boldsymbol{\beta} + \text{offset}. \tag{2}$$

Thus, the offset behaves like an additional regressor whose coefficient is not estimated but fixed, e.g. to 1. A prominent example for offsets in GLMs is the modeling of rates in Poisson regression, where offset $= 1 \cdot \log(\text{exposure})$.

GLM trees

Tree algorithms generally split the data recursively into disjoint subgroups (also called nodes) starting from the so-called root node containing all data and employing certain split points in the so-called split variables. In case of GLM trees, the idea is to (1) *estimate* the parameters in a GLM using the current sample (starting with the full data set), (2) *assess* whether the parameters are stable over the split variables considered, (3) *split* the sample along the variable associated with the highest parameter instability, (4) *repeat* the previous steps recursively until some stopping criterion is met (e.g., with respect to the size of the sample or the instability of the parameters). Various algorithms have been suggested that can be employed for such GLM-based recursive partitioning, including GUIDE (Loh 2002), CTree (Hothorn, Hornik, and Zeileis 2006), or MOB (Zeileis et al 2008) where the latter is used subsequently and explained in more detail in Section 2.2.1.

Figure 1 shows an example tree structure that could be found by a GLM tree with

$$\boldsymbol{\beta}(\mathbf{z}) = \begin{cases} \boldsymbol{\beta}_1 & \text{if } \; z_1 \leq 0 \\ \boldsymbol{\beta}_2 & \text{if } \; (z_1 > 0) \; \wedge \; (z_2 \leq 0) \\ \boldsymbol{\beta}_3 & \text{if } \; (z_1 > 0) \; \wedge \; (z_2 > 0). \end{cases} \tag{3}$$

The parameters $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\beta}_3$ can be estimated by three separate models for the three subgroups or by using interaction terms as in Equation 1 ($I(\text{subgroup}_1) = I(z_1 \leq 0)$ etc.). To make the role of the split variables more explicit we from now on write $\mathbf{x}^\top \boldsymbol{\beta}(\mathbf{z})$ instead of $\tilde{\mathbf{x}}^\top \tilde{\boldsymbol{\beta}}$. $\boldsymbol{\beta}(\mathbf{z})$ is the interaction effect between covariates $\mathbf{x}$ and the subgroups defined by the split variables $\mathbf{z}$.

2.2. Extension: PALM trees

GLM trees assume that all parameters are subgroup specific. This does not necessarily have to be the case. PALM trees address this issue and offer a compromise between GLM trees and GLMs by having one part in which the parameters depend on subgroups (these are again denoted by $\boldsymbol{\beta}(\mathbf{z})$) and another part in which the parameters are the same for all subjects/subgroups (denoted by $\boldsymbol{\gamma}$).

Going from GLMs via GLM trees to PALM trees can be viewed as an evolutionary process where one method evolves from the other. The goal of all three is to appropriately estimate the effect of covariates $\mathbf{x}$ on an outcome y. The main difference between the three methods is the structure of the linear predictor. While the effects $\boldsymbol{\beta}$ are linear in a GLM, the effects $\boldsymbol{\beta}(\mathbf{z})$ are linear and constant within each subgroup but vary between subgroups, i.e. are subgroup-wise linear. A PALM tree contains globally *fixed* linear effects $\boldsymbol{\gamma}$ for some covariates $\mathbf{x}_F$ and subgroup-wise *varying* linear effects $\boldsymbol{\beta}(\mathbf{z})$ for other covariates $\mathbf{x}_V$. Mathematically this can be expressed as follows:

$$\text{GLM} \quad g(\mu) = \mathbf{x}^\top \boldsymbol{\beta} \tag{4}$$

$$\text{GLM tree} \quad g(\mu) = \mathbf{x}^\top \boldsymbol{\beta}(\mathbf{z}) \tag{5}$$

$$\text{PALM tree} \quad g(\mu) = \mathbf{x}_V^\top \boldsymbol{\beta}(\mathbf{z}) + \mathbf{x}_F^\top \boldsymbol{\gamma}. \tag{6}$$

In PALM trees the variables $\mathbf{x}_F$ with a global effect $\boldsymbol{\gamma}$ have to be defined a priori. Usually $\mathbf{x}_V$ and $\mathbf{x}_F$ and $\mathbf{z}$ do not overlap although this is, in principle, possible. Note that if the subgroup structure were known, models 5 and 6 could both be estimated as GLMs. Only the fact that it is unknown and has to be detected makes GLM trees and PALM trees necessary. Also, if the global parameter vector $\boldsymbol{\gamma}$ were known, model 6 could be estimated as GLM tree with $\mathbf{x}_F^\top \boldsymbol{\gamma}$ as offset (as in equation 2). These connections between the methods are leveraged in the PALM tree algorithm.

Algorithm

We now describe the detailed GLM tree and PALM tree algorithms, starting with GLM trees as the PALM tree algorithm uses GLM trees in the estimation process. The GLM tree algorithm is not new and has been explained in depth by Zeileis et al (2008). The following description of the algorithm focuses on the parts that are necessary in order to demonstrate the full concept of the PALM tree algorithm. Note that to notationally distinguish the parameters

in the subgroups (e.g. parameter vector in first subgroup $\boldsymbol{\beta}_1$) from parameters in the models (e.g. first model parameter $\beta_{(1)}$) we use parentheses. GLM trees are grown as follows, starting with the root node containing all observations:

1. Compute model (4), or equivalently model (5) with a single subgroup ($\boldsymbol{\beta}(\mathbf{z}) = \boldsymbol{\beta}$), in the given node.

2. Test for instability in the model parameters with respect to each of the possible subgroup defining variables $Z_1, \ldots, Z_J$:

   - Compute the score contributions

   $$s_{(k)}\left((y, \mathbf{x})_i, \hat{\boldsymbol{\beta}}\right) = \left.\frac{\partial l\left((y, \mathbf{x})_i, \boldsymbol{\beta}\right)}{\partial \beta_{(k)}}\right|_{\hat{\boldsymbol{\beta}}}$$

   as the partial derivatives of the log-likelihood contributions of each observation i ($i = 1, \ldots, n$) with respect to the model parameters $\beta_{(1)}, \ldots, \beta_{(K)}$ evaluated at the estimated parameters $\hat{\boldsymbol{\beta}} = (\hat{\beta}_{(1)}, \ldots, \hat{\beta}_{(K)})^\top$.

   - Test if the scores fluctuate randomly around zero for each variable Z_j ($j = 1, \ldots, J$) , i.e.

   $$H_0^{\beta_{(k)}, j}: \quad S_{(k)}\left((Y, \mathbf{X}), \hat{\boldsymbol{\beta}}\right) \perp Z_j$$

   using M-fluctuation tests (Zeileis and Hornik 2007).

3. If the overall test is significant (usually multiplicity adjustment using Bonferroni correction is used here), choose variable Z_j corresponding to the lowest p-value as the split variable. In the following, we will use 5% as the global significance level.

4. Choose as split point the point in the split variable which maximizes the sum of likelihoods in the emerging subgroups.

5. Iterate steps 1 to 4 until $H_0^{\beta_{(k)}, j}$ $\forall k, j$ cannot be rejected or some other stop criterion (e.g. minimum subgroup size is reached) is fulfilled.

The resulting groups differ with respect to at least one of the model parameters $\boldsymbol{\beta}$. In practice, however, all parameters vary slightly between subgroups due to the refitting of the model in each node, i.e. for each group of observed subjects. If in reality some covariates influence the response linearly (for all observations), this leads to an overly complex model. The PALM tree algorithm eliminates this downside by introducing the possibility to build models where some parameters are kept stable across subgroups. This is achieved by starting the estimation of model (6) with a single subgroup, i.e. $\boldsymbol{\beta}(\mathbf{z}) = \boldsymbol{\beta}$, and then iterating the tree growing process between

(a) estimating $\boldsymbol{\gamma}$ for a given tree structure and

(b) estimating the tree structure for a given $\hat{\boldsymbol{\gamma}}$ (steps 1.-5.).

In (a) we estimate the full model (6) for the known subgroup $\times$ covariate ($\mathbf{x}_V$) interactions (as in equation 1) and get estimates for $\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\gamma}$. In (b) we treat the estimated $\hat{\boldsymbol{\gamma}}$ as fixed and include $\mathbf{x}_V^\top \hat{\boldsymbol{\gamma}}$ in the model as an offset. By preventing $\boldsymbol{\gamma}$ from being estimated, we exclude it

from the score function and can grow a standard GLM tree (as in steps 1.-5.) for the remaining parameters. At the same time we want to account for the effects of $\mathbf{x}_V$ which is obtained by including $\mathbf{x}_V^\top \hat{\boldsymbol{\gamma}}$ as offset. The iterative process stops when no (or very little) improvement in terms of log-likelihood can be achieved (typically when the tree structure does not change anymore). Iterating between (a) and (b) simplifies estimation by only having one unknown: either $\boldsymbol{\gamma}$ or the tree structure. $\boldsymbol{\beta}(\mathbf{z})$ is estimated in both steps: In (a) by estimating the model with the known subgroup $\times$ covariate interactions, and in (b) by estimating a separate model for each subgroup.

PALM trees inherit many of their theoretical properties from the methods used as building blocks (model-based trees and parametric models), provided that the model is well specified: Given that the group structure is correctly detected by the tree, the (G)LM can consistently estimate all coefficients (grouped and global). Conversely, given that the global coefficients are estimated consistently, the (G)LM tree uses a group detection based on locally consistent tests (Zeileis and Hornik 2007) and the usual locally optimal greedy forward selection in recursive partitioning (see e.g. Breiman, Friedman, Stone, and Olshen 1984). To the best of our knowledge, there is no formal proof that alternating between (a) and (b) will converge to an "optimal" solution so that the strengths of both components are guaranteed to be effective. However, our simulation results (see Section 3 and Appendix A) show that PALM trees typically converge quickly and reliably. This was also found for RE-EM trees (Sela and Simonoff 2012). While there is no guarantee that this is always the case, we have not experienced any convergence issues thus far.

2.3. Special application: Treatment effects

One common application of model-based trees is for subgroup analyses in clinical trials (Lipkovich, Dmitrienko, and D'Agostino 2016; Seibold et al 2016; Doove et al 2014). In the simplest case one is interested in a treatment effect of a new treatment versus standard of care or no treatment, i.e. $\mathbf{x}$ or $\mathbf{x}_V = (1, x_A)$ with $x_{Ai} = I(\text{patient } i \text{ received new treatment})$. In this setting one differentiates between prognostic and predictive factors (Italiano 2011). Prognostic factors are patient characteristics (measured before treatment start) which directly impact the response, e.g. a health score. Predictive factors are patient characteristics which impact the efficacy of the treatment. In the PALM tree framework, predictive factors should be included in the split variables $\mathbf{z}$ and prognostic factors, if known in advance, can be included in $\mathbf{x}_F$. In fact, prognostic factors are often known in advance based on previous research about the disease.

In subgroup analyses for treatment effects the term optimal treatment regime is commonly mentioned. An optimal treatment regime is a rule which indicates which treatment is better in which subgroup. Treatment regimes only check the sign of the treatment effect in each subgroup. If they differ between subgroups, the treatment effects are called qualitative; if one treatment is better than the other in all subgroups, they are called quantitative. As this application is very common, the remainder of this manuscript will deal with scenarios where the partitionable parameters are the intercept and the effect of a binary covariate.

2.4. Comparison to other approaches

GLMM trees (Fokkema et al 2017) are closely related to PALM trees, as the algorithm also builds on the GLM tree algorithm and like PALM tree keeps parts of the model stable. The

major difference is the fact that GLMM trees focus, as the name says, on generalised mixed effects models and the part that is being kept stable across subgroups are the random effects.

STIMA (Dusseldorp et al 2010) is a tree algorithm where the first split is made in an a priori specified variable, which in the treatment case is the treatment indicator. All further splits are found by an exhaustive search and finally a cross-validation based pruning procedure is run to find the optimal tree. STIMA is similar to PALM tree in the sense that it starts off with a main effects model and new splits are selected based on a measure of variance-accounted-for. The main effects of the model are kept stable across groups and additional effects are added to the model based on the tree structure. A very similar approach is called partially linear tree-based regression model (PLTR, Chen, Yu, Hsing, and Therneau 2007; Mbogning and Toussile 2015), which was initially invented to analyse gene-gene and gene-environment effects.

The approach by Zhang et al (2012) aims to estimate optimal treatment regimes and is only used in the treatment effect application. In the following we will use the term OTR (optimal treatment regimes) for this method. OTR is not as closely related to PALM tree as the previously mentioned methods, but has shown good performance in settings in which PALM trees are appropriate (Sies and Van Mechelen 2017). OTR does not target estimating the treatment effect itself but targets learning which treatment is superior for certain groups of patients. OTR starts off with the so-called outcome model, which includes main effects and treatment $\times$ patient characteristics interactions. After estimating the model the algorithm proceeds as follows:

1. For all patients in the training data predict the response under treatment $\hat{\mu}_1$ and under control $\hat{\mu}_0$ from the outcome model. Determine the difference $\hat{\mu}_1 - \hat{\mu}_0$ between the two.
2. Compute a classification algorithm using $I(\hat{\mu}_1 - \hat{\mu}_0 > 0)$ as response and $|\hat{\mu}_1 - \hat{\mu}_0|$ as weights.

Any classification method that can deal with (non-integer) weights could be used in step 2.

For further tree-based approaches that allow doing analyses similar to model-based trees see Doove et al (2014).

3. Simulation study

We compare the performance of PALM trees, LM trees, the trees grown based on the algorithm proposed by Zhang et al (2012) (OTR) and STIMA in the treatment effect setting. We chose OTR as competitor because it showed good perfomance in scenarios where PALM trees should perform well (Sies and Van Mechelen 2017) and we chose STIMA because it is a natural competitor due to the similarity of the resulting model. Note that while the setup of the simulation study is motivated by treatment effect studies, the insights are of broader interest due to its general structure. The aim is to evaluate the methods with respect to (1) finding the correct subgroups (Section 3.1), (2) not splitting when there are no subgroups (Section 3.2), (3) finding the optimal treatment regime (Section 3.3), and (4) correctly estimating the treatment effect (Section 3.4). Note that evaluations (1) and (2) are connected in the sense that they both evaluate the ability to find the correct subgroups. Furthermore, (3) and (4) are connected in the sense that both evaluate the ability to give good treatment recommendations.

Simulation variable	Default	Variation	# Values
Difference in treatment effects Δ_β	0.5	0.1–1.5	8
Number of observations n	300	100–900	5
Qualitative treatment × subgroup interaction	Yes	Yes/No	2
Number of patient characteristics m	30	10–70	4
Number of predictive factors p	2	1–4, 0	4, 1
Number of prognostic factors q	2	1–4	4

Table 1: Simulation settings. For each scenario one simulation variable is varied and the rest are kept to the standard value. The value $p = 0$ is only used for the assessment of the type 1 error rate (Section 3.2).

We simulate a binary variable (treatment indicator) X_A which is either 1 or 0, each with probability 0.5, and m correlated variables (patient characteristics)

$$\mathbf{Z} \sim \mathcal{N}_m(\mathbf{0}, \mathbf{\Sigma}) \tag{7}$$

with

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0.2 & \cdots & 0.2 \\ 0.2 & 1 & \cdots & 0.2 \\ \vdots & \vdots & \ddots & \vdots \\ 0.2 & 0.2 & \cdots & 1 \end{pmatrix}. \tag{8}$$

We define the first p variables $Z_1, \ldots, Z_p$ to be the true predictive factors, i.e. the patient characteristics that actually interact with the treatment and thus pose relevant split variables. The cutpoint is always at $Z_j = 0$ and the subsequent split is always in the subgroup with $Z_j > 0$, i.e. on the right side of the tree when visualised as in Figure 1. We define the consecutive q variables $X_F = (Z_{p+1}, \ldots, Z_{p+q})$ to be the true and known prognostic factors. All further patient characteristics $Z_{p+q+1}, \ldots, Z_m$ are noise variables. We simulate the outcome variable Y with

$$Y = X_A \beta(\mathbf{Z}) + X_F \boldsymbol{\gamma} + U \tag{9}$$

where $U \sim \mathcal{N}(0, 1.5)$ is the error term.

The effect of the prognostic factors is set to $\boldsymbol{\gamma} = \mathbf{1}$. The treatment effect $\beta(\mathbf{Z})$ follows a tree structure, which is visualised in Figure 1 for the scenarios with $p = 2$. The mathematical representation is as in Equation (3) with a fixed difference between the effects in the subgroups Δ_β. We define a default simulation scenario, which is shown in the second column of Table 1. In this default scenario $\Delta_\beta = 0.5$ and

$$\beta(\mathbf{Z}) = \begin{cases} -0.375 & = \beta_1 & \text{if } Z_1 \leq 0 \\ 0.125 & = \beta_2 = \beta_1 + \Delta_\beta & \text{if } Z_1 > 0 \ \wedge \ Z_2 \leq 0 \\ 0.625 & = \beta_3 = \beta_2 + \Delta_\beta & \text{if } Z_1 > 0 \ \wedge \ Z_2 > 0. \end{cases} \tag{10}$$

To obtain a diverse set of simulation scenarios which are comparable, we fix all but one of the simulation variables to the default. The range of variation of each simulation variable is

given in the third column of Table 1 alongside the number of equidistant values considered (# Values). From this we get all necessary information about the simulation, e.g. q takes 4 different values $1, 2, 3, 4$. For each distinct simulation setting we simulate 150 data sets. Note that just for the assessment of the type 1 error rate (Section 3.2) the number of predictive factors is set to zero. For the simulation scenarios where $p \neq 2$ and thus less/more than three true subgroups exist, $\beta(\mathbf{Z})$ follows the same logic as in Equation (10), i.e. $\beta_b = \beta_{b-1} + \Delta_\beta$ for $b = 2, \ldots, (p+1)$. The value of β_1 depends on whether the first split is qualitative or not and on Δ_β. If the first split is not qualitative then $\beta(1) = 0.5$. If the first split is qualitative $\beta(1) = -3/4 \cdot \Delta_\beta$. This also means that any consecutive splits after the first are quantitative. This simulation study is limited due to the fact that we only change one simulation variable at a time. Section A in the Appendix shows selected results from a full factorial simulation study. Using the simulated data we compare the following methods:

PALM tree with $\mathbf{x}_V = (\mathbf{1}, \mathbf{x}_A)$ and $\mathbf{x}_F = (z_{p+1}, \ldots, z_{p+q})$. The only way we could have specified this algorithm better for the given data generating process would have been to add the intercept to $\mathbf{x}_F$, but in real application one would usually allow the intercept to vary to account for unknown prognostic factors contained in $\mathbf{z}$.

LM tree 1 with $\mathbf{x} = (\mathbf{1}, \mathbf{x}_A)$. This algorithm is of interest to see how well a misspecified model-based tree behaves. LM tree 1 has to approximate $\mathbf{x}_F^\top \boldsymbol{\gamma}$ using step functions and thus cannot give good results in terms of most measures used below. However, we are interested in how well it can do in terms of estimating the correct treatment regime.

LM tree 2 with $\mathbf{x} = (\mathbf{1}, \mathbf{x}_A, \mathbf{x}_F)$. This tree is expected to behave better than LM tree 1, since it contains the correct covariates in the model, but worse than PALM tree since it may split with respect to instabilities in the parameters for $\mathbf{x}_F$ plus it is overly complex due to the fitting of separate $\mathbf{x}_F$-parameters in each subgroup.

OTR with outcome model $g(\boldsymbol{\mu}) = (\mathbf{1}, \mathbf{x}_A, \mathbf{x}_F)^\top \boldsymbol{\gamma} + (\mathbf{x}_A : \mathbf{z})^\top \boldsymbol{\beta}$ (with $\mathbf{x}_A : \mathbf{z}$ interaction between $\mathbf{x}_A$ and $\mathbf{z}$) and pruned CARTs (Classification and Regression Trees, Breiman et al 1984) as classification method. OTR was invented to find optimal treatment regimes and thus is expected to be good at finding the right treatment. OTR is not intended to find quantitative interactions and thus can not be good at this.

STIMA with a forced first split in the treatment and the maximum number of splits fixed to six.

3.1. Are the correct subgroups found?

To investigate whether the correct subgroups are captured by the different methods, we looked at the number of subgroups found as well as the adjusted Rand index (ARI, Hubert and Arabie 1985; Milligan and Cooper 1986). The ARI measures how well the retrieved subgroups fit with the true underlying subgroups. If the subgroups found are similar to the true subgroups the ARI will have a value up to 1. If the subgroups are only as good as a random group assignment the ARI is 0. If there is systematic missclassification, the ARI can also be negative.

The first row of Figure 2 shows the mean number of selected subgroups over the 150 simulated data sets and their corresponding trees for differing *distances between treatment effects* Δ_β and differing *numbers of observations* n. This means we are looking at the case where all

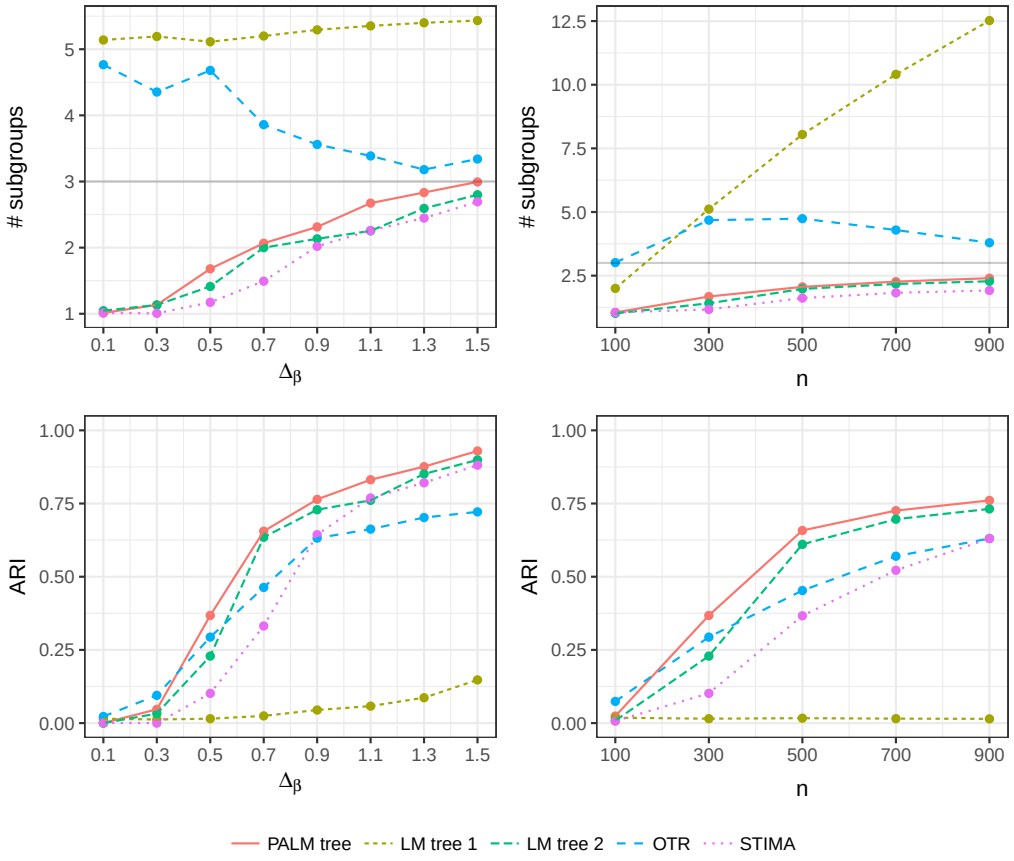


Figure 2: Mean number of subgroups and mean ARI for varying Δ_β and number of observations (Question 3.1).

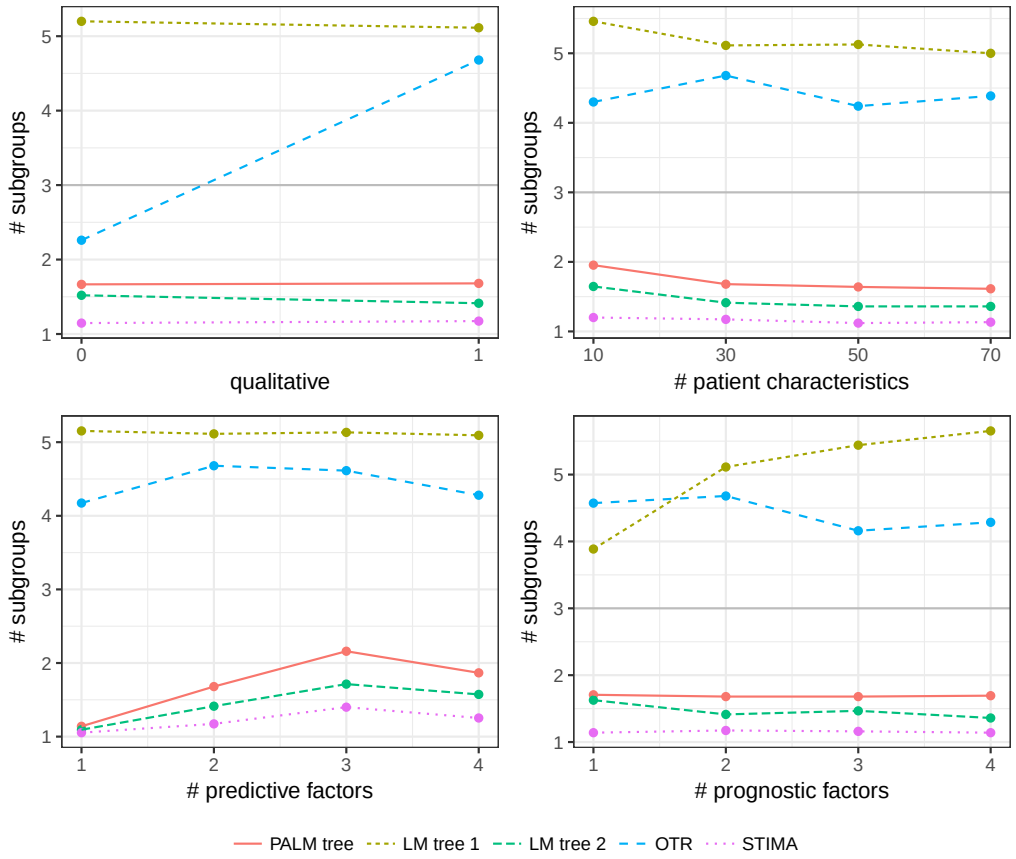


Figure 3: Mean number of subgroups for varying types of subgroups (quantitative/qualitative), number of patient characteristics, predictive factors and prognostic factors (Question 3.1).

variables are kept at the standard value except Δ_β or n respectively. The second row shows the corresponding ARI. The similarity between the PALM tree and LM tree 2 algorithms is obvious. For both the number of subgroups and the ARI the results are very similar, although PALM tree is slightly better. Both algorithms get steadily closer to the optimal solution with increasing Δ_β as well as with increasing number of observations. LM tree 1 performs badly because it approximates the linear relation between the prognostic factors and the response with splits in the data. This is also the reason why with increasing n the number of subgroups increases. This effect muffles the grouping with respect to the treatment effect, even if it gets less with increasing Δ_β. The number of subgroups found for OTR is on average greater than the actual number of subgroups (3 for the given scenarios in Figure 2). The variability of the number of subgroups for OTR is very high (with a maximum of 20 subgroups). The true subgroups are not captured as well as with PALM tree and LM tree 2. The ARI for OTR is lower than the ARI of PALM tree and LM tree 2 except for very low values of Δ_β and n, which can be explained by the fact that the model-based trees use statistical significance tests and CART does not. Even though the pattern of STIMA in terms of the average number of subgroups appears similar to PALM tree and LM tree 2, on average the ARI is considerably lower, except for very large differences in treatment effects (Δ_β).

Figure 3 shows the mean number of subgroups for the remaining simulation scenarios. The model-based trees and STIMA are not affected by the *type of subgroup*. OTR, however, is designed to find only qualitative subgroups and thus on average finds fewer groups when there are only quantitatively differing subgroups. For increasing *number of patient characteristics*, the model-based trees become more conservative and find slightly less subgroups, which is due to the correction for multiple testing (Bonferroni correction). OTR and STIMA do not change much in terms of average number of subgroups when the number of patient characteristics increases. With increasing *number of predictive factors* the number of subgroups should increase. The true number of subgroups is always the number of predictive factors + 1. The lower left panel of Figure 3 shows that this is not the case for any of the algorithms. The reason for this is the way of how we simulated the data. With an increasing number of predictive factors the subgroups get smaller and thus there is less power to find splits. The only algorithm that is strongly affected by the *number of prognostic factors* is LM tree 1, which corresponds to the fact that there are more linear terms to approximate through the tree structure.

3.2. How often are subgroups found even though there are none?

To investigate the type 1 error rate, i.e. the probability that subgroups are found even though there are none, we simulated data as above, but with no predictive factors. This means the treatment effect is the same for all patients. Figure 4 shows the behaviour of the methods with changing *number of observations*. LM tree 1 and OTR have a constant value of 1 here and are not visualised. Since LM tree 1 finds subgroups that have to do with the prognostic factors the "bad" performace exists by design. PALM tree is close to the expected 5% significance level, as is LM tree 2. STIMA goes down to 0% for 700 and 900 observations.

3.3. Is the correct treatment predicted to be better?

The next measure we wanted to look at is the proportion of patients for which the better treatment is correctly identified. This is what OTR was designed to be good at and especially

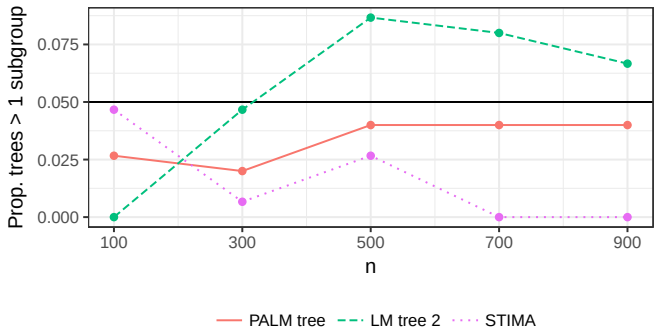


Figure 4: Proportion of trees with more than one subgroup for varying number of observations (Question 3.2). Black line at 0.05.

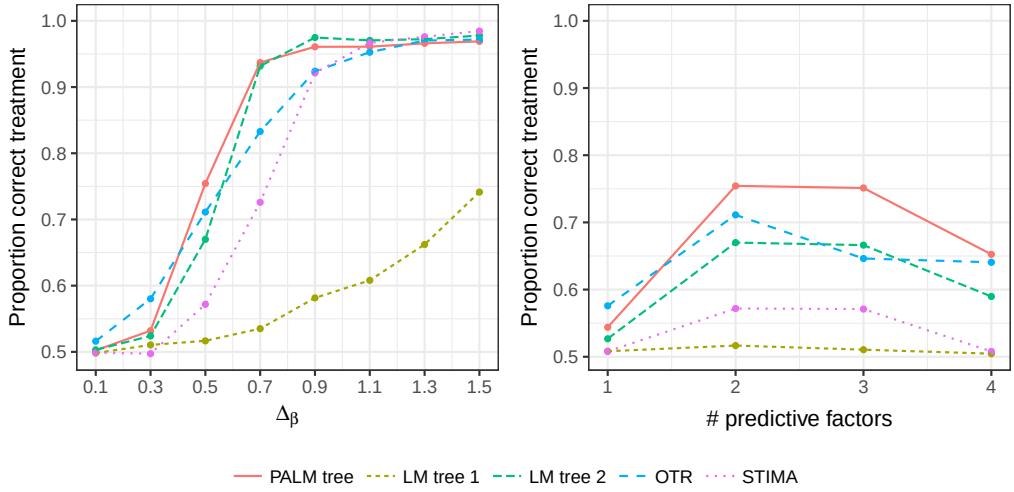


Figure 5: Proportion of observations in all trees where better treatment is correctly identified (Question 3.3).

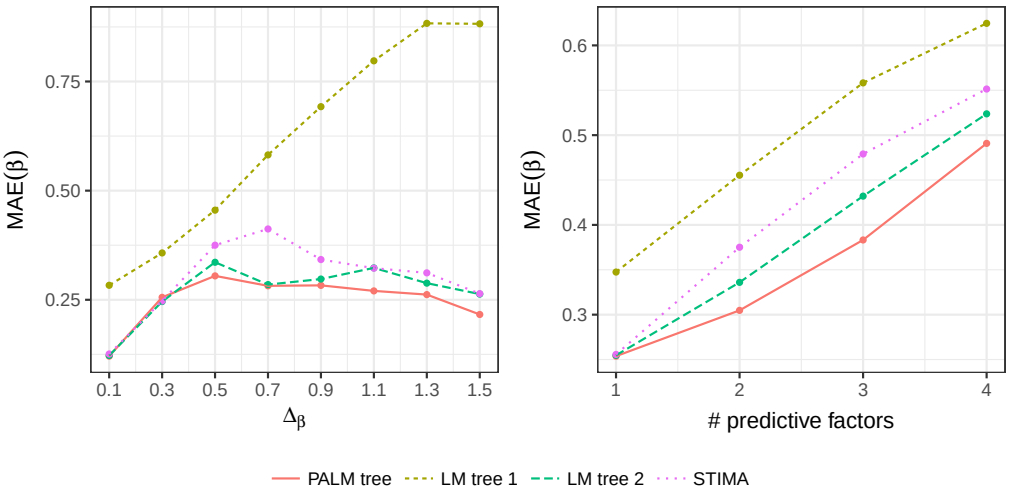


Figure 6: Mean absolute difference between true and estimated treatment effect (mean absolute error, MAE; Question 3.4).

due to the way we simulated data (with a simple interaction) OTR can be expected to perform well. Figure 5 shows the proportion of patients for which the better treatment is correctly identified for the scenarios with varying difference between treatment effects Δ_β and varying number of predictive factors. When the *difference between treatment effects* Δ_β is small it is difficult for all methods to predict the correct treatment regime. For $\Delta_\beta = 0.1$ it is close to random guessing. With increasing Δ_β all methods get better. The performance of PALM tree, LM tree 2, OTR and STIMA is similar. The four methods also behave similarly with a changing *number of predictive factors*. The treatment regime prediction is globally worst on average when there is one predictive factor. This results from the fact that often no split is found in this case (see Figure 3). In cases where the methods decide not to split at all, this leads by simulation design to a proportion of 50% correctly-defined treatment regimes. The proportion of patients for which the correct treatment is predicted to be the better treatment improves in cases of two or three predictive factors and gets worse with four predictive factors. With more complex and smaller subgroups it becomes more difficult for the algorithms to retrieve the correct subgroup structure and to estimate the treatment effect. Note, however, that shape of the shape of the curves in the right panel of Figure 5 is very specific for the simulation settings here. Figure 9 shows the results for other scenarios. For example, for $\Delta_\beta = 1.5$ and 300 observations in a setting with qualitative treatment differences, the best performace of PALM tree is with only one predictive factor and decreases from there. The performance of all algorithms is well in quantitative settings. OTR is the only algorithm that goes down to only 80% correctly defined treatment regimes in settings with 100 observations.

3.4. How good is the treatment effect estimate?

Estimating or even predicting the correct treatment effect is the most essential part of subgroup analysis. Even if one treatment is better than the other, clinicians need to know if the

difference is relevant. The evaluation of the treatment effect estimate can only be done for the model-based recursive partitioning methods and STIMA since OTR is only designed to produce binary decision rules. The measure used to evaluate the treatment effect estimate is the mean absolute difference between true and estimated treatment effect (mean absolute error, MAE). Figure 6 shows the MAE for the scenarios of varying Δ_β and varying number of predictive factors. The error is smallest for all three methods when the *difference in treatment effect* is lowest ($\Delta_\beta = 0.1$), because even if the chosen subgroups are wrong, the estimated treatment effect will likely be close to the true and very similar treatment effects. In this sense it is not a disadvantage that PALM tree, LM tree 2 and STIMA often do not split into subgroups at all. In fact, it may even be an advantage, as the treatment effect estimate is then calculated based on a larger data set and is less affected by random variability. The effect of the small treatment difference gets less as the difference increases. However, as the it increases, finding the correct subgroups becomes easier and the error decreases. At the same time finding the correct subgroups becomes easier and slowly the error decreases again for PALM tree, LM tree 2 and STIMA. For this effect to be visible for LM tree 1, one would have to have larger treatment effects, fewer prognostic factors and/or more observations, given the large effect of the prognostic factor (see Figure 11 in the Appedix). With an increasing *number of predictive factors* the mean absolute error in treatment effect increases. The shape of the curve in Figure 6 looks very different to the one in Figure 5, even though they address similar questions, but the more true predictive factors exist in the given simulation scenario the harder it is for the methods to predict the treatment effect. This suggests that simply knowing the more effective treatment does not tell the whole story. This is supported across simulation scenarios (compare Figures 9 and 10).

4. Illustration: Treatment differences in mathematics exam

The Mathematics 101 course for first-year business and economics students at Universität Innsbruck gives an introduction to mathematical analysis, linear algebra, financial mathematics, and probability calculus. Students are assessed by biweekly online tests during the semester and a written exam at the end. The exam consists of 13 single-choice questions with 5 answer alternatives, one of which is correct. Students who answer more than 60 percent of the questions correctly pass the course. The percentage of successful online tests captures math ability of the students and is a known predictor for success in the final exam.

The data contains the exam results of 729 students (out of 941 who originally registered for the course) for the fall semester in 2014/15. Due to limited availability of seats in the exam room, the students were asked to select a group, where the first group wrote the exam in the morning and the second group right after the first group finished. The two groups received slightly different questions on the same topics covering the scope of the course. We are interested in whether the exam is fair in the sense that it is on average equally hard or difficult for the two groups. In other words we want to find out whether there is a "treatment effect" with the different selection of exam questions in the two groups corresponding to the "treatments". As a first rather naive check we consider a simple one-way regression model for the percentage of correct answers by group, as reported in the first column of Table 2. This yields an expected percentage of 57.6 for a student in group 1 and a difference of 2.33 percentage points for students in group 2. Thus, the model finds only a small drop in the percentage of correctly solved answers and the corresponding confidence interval includes a

zero change.

However, in this first model we have neglected the influence of the students' ability which is particularly relevant here because the students could freely choose their exam group. Therefore, there might have been self-selection of more (or less) able students into the first (or second) group. To account for such ability effects in the model we include the percentage of points from the previous online tests that captures the students' ability and preparation. As shown in the second column of Table 2 this variable is indeed strongly associated with the exam results, where one additional percentage point in the online tests leads to additional 0.86 expected percentage points in the written exam. More importantly, the group effect increases to 4.37 and the corresponding confidence interval does not include zero anymore. Despite the increase in the group effect, the absolute size of the group difference is still moderate corresponding to about half an exercise out of 13.

To explore the size of the treatment effect for the group differences further, we consider the possibility that this may vary across subgroups of students. Known student characteristics that may lead to such subgroups here are gender, the number of semesters the student has already been studying, the number of times the student has already attempted the exam, the type of study (three year bachelor program vs. four year diploma program) and also the ability/preparation as captured by percentage of successful exercises in the online tests. Since the test results in the online tests during the semester are known to have an important direct effect on the performance in the exam, the test parameter is included in the PALM tree. Figure 7 shows the resulting PALM tree with the segmented local group effect while adjusting for a global online tests effect. The strongest parameter instability is associated with the number of attempts and the group of students in the first attempt are split a second time by the percentage from the online tests. Two of the resulting subgroups (node 3 and 5) exhibit only very small group differences but in node 4 the second group obtained clearly a lower response percentage. This node is the smallest subgroup found and encompasses the highly able students taking the course for the first time. For this subsample the treatment effect is about 14 percentage points, which means that the students in the second batch solved about two exercises less than those in the first batch.

Overall this clearly conveys the strength of the PALM tree method: Especially in situations where the coefficient of interest is modest in a main-effects model and where further covariates are available whose influence on the main model parameters is not obvious, the PALM tree is an attractive option to globally control for certain variables while searching for local effects in others. Note, however, that due to the forward selection of models/effects the resulting confidence intervals in the terminal nodes (Table 2 and Figure 7) should not be used for inference but interpreted as a measure of variability.

5. Discussion

Model-based trees are effective tools to identify subgroups in data which differ in terms of model parameters. PALM trees are special model-based trees where some parameters can be fixed globally for the entire sample and do not depend on the subgroup structure. Our simulation study has shown that in cases where there are such specified factors with a direct effect on the outcome, PALM trees reliably detect the correct subgroups while at the same time having a low probability of detecting subgroups when there are none. STIMA is a

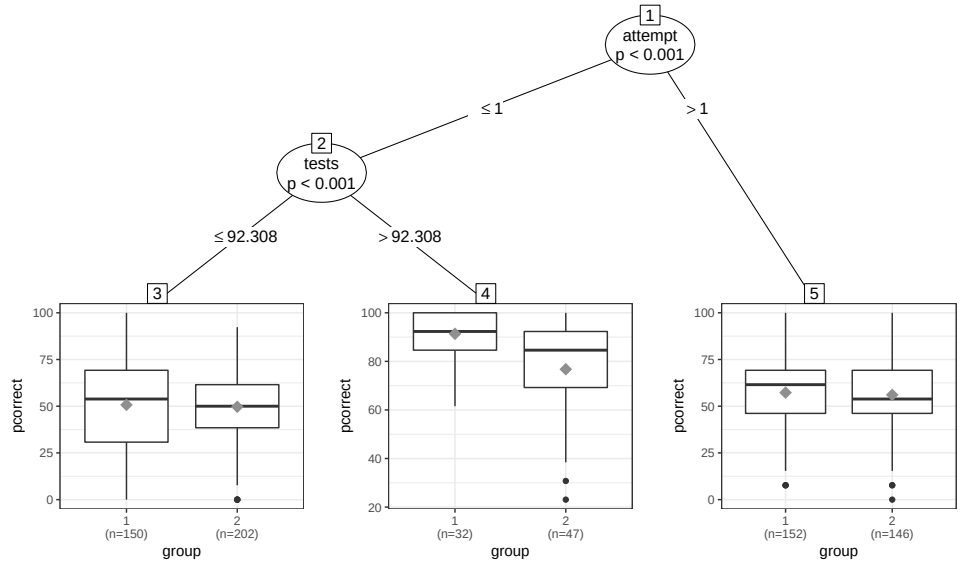
	Linear model 1	Linear model 2	PALM tree
(Intercept)	57.60 [55.12, 60.08]	−5.85 [−13.52, 1.83]	
node3:(Intercept)			−7.09 [−16.15, 1.97]
node4:(Intercept)			13.98 [0.82, 27.14]
node5:(Intercept)			2.33 [−6.32, 10.99]
group2	−2.33 [−5.70, 1.03]	−4.37 [−7.23, −1.50]	
node3:group2			−3.00 [−6.97, 0.98]
node4:group2			−14.49 [−22.92, −6.07]
node5:group2			−1.70 [−5.97, 2.56]
tests		0.86 [0.76, 0.95]	0.79 [0.67, 0.90]

Table 2: Three models for the mathematics exam data. The response variable is the percentage of correctly solved exercises and the main covariat of interest are the treatment differences between the first and second exam group. Confidence intervals are given in brackets.

flexible and well performing competitor of model-based trees. The most important downside of STIMA is that it is very slow with in some instances single trees taking hours to compute (see Appendix B). Moreover, it has to be taken with a grain of salt that the R package "stima" is not actively maintained on the Comprehensive R Archive Network. Although optimal treatment regimes (OTR) perform comparably to PALM trees in terms of detecting the best treatment option in the given simulation study, PALM trees are typically better at recovering a parsimonious tree capturing the underlying subgroup structure. This makes PALM tree results easier to interpret and to communicate to practitioners, which we believe is an important advantage in many applications. Moreover, the simulation study clearly showed the effect of misspecifications in global vs. local effects in PALM trees. While it is important to correctly identify the variables with *additive* effects (LM tree 1 vs. LM tree 2 or PALM tree), it is not so important to correctly identify whether these additive effects are *global or local* (LM tree 2 vs. PALM tree). However, by reducing the number of tests in the split procedure and focusing only on certain relevant model parameters, some power and efficiency can be gained from selecting a suitable PALM tree.

PALM trees allow exploring and questioning results of (generalised) linear models. The PALM tree analysis of the Mathematics 101 exam showed that a linear model regressing the percentage points of correct anwers on the group and earlier test results is too simple. Only for a relatively small subgroup of students who attempted the exam for the first time and who showed good performance during the semester it did make a difference whether they attempted the exam in the first or second group.

Although large parts of this manuscript focus on subgroup analyses in clinical trials, PALM



	Estimate	2.5%	97.5%
(Intercept)	−7.09	−10.40	−3.78
group2	−3.00	−7.37	1.37
tests	0.79	0.67	0.90

	Estimate	2.5%	97.5%
(Intercept)	13.98	8.01	19.95
group2	−14.49	−22.24	−6.75
tests	0.79	0.67	0.90

	Estimate	2.5%	97.5%
(Intercept)	2.33	−0.33	4.99
group2	−1.70	−5.51	2.10
tests	0.79	0.67	0.90

Figure 7: PALM tree for the percentage of correct answers explained by group differences while globally adjusting for ability (i.e., percentage of points obtained in previous online tests).

trees can also be applied in a wide range of other applications as well – e.g., in the social sciences as shown in the mathematics exam application case study.

Computational details

Open-source implementations of the model-based tree algorithms LM tree and GLM tree are available in the **partykit** package (Hothorn and Zeileis 2015, functions `lmtree()` and `glmtree()`). The PALM tree algorithm is available in the **palmtree** package (Seibold, Hothorn, and Zeileis 2017, function `palmtree()`). OTR is available in package DynTxRegime (Holloway, Laber, Linn, Zhang, Davidian, and Tsiatis 2015). The STIMA implementation has been archived on CRAN but can still be downloaded from `https://cran.r-project.org/src/contrib/Archive/stima/`. Simulations were conducted using the batchtools package (Lang, Bischl, and Surmann 2017).

The manuscript including simulation study and application can be reproduced using the supplementary online material.

Acknowledgements

We thank Andrea Farnham for improving the language. We are thankful to the Swiss National Fund for funding this project with grants 205321_163456 and IZSEZ0_177091 and mobility grant 205321_163456/2.

References

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and Regression Trees. Wadsworth

Chen J, Yu K, Hsing A, Therneau TM (2007) A partially linear tree-based regression model for assessing complex joint gene-gene and gene-environment effects. Genetic Epidemiology 31(3):238–251, doi:10.1002/gepi.20205

Doove LL, Dusseldorp E, Van Deun K, Van Mechelen I (2014) A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment-subgroup interactions. Advances in Data Analysis and Classification 8(4):403–425, doi:10.1007/s11634-013-0159-x

Dusseldorp E, Conversano C, Van Os BJ (2010) Combining an additive and tree-based regression model simultaneously: Stima. Journal of Computational and Graphical Statistics 19(3):514–530, doi:10.1198/jcgs.2010.06089

Fokkema M, Smits N, Zeileis A, Hothorn T, Kelderman H (2017) Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. Behavior Research Methods In press

Hajjem A, Bellavance F, Larocque D (2011) Mixed effects regression trees for clustered data. Statistics & Probability Letters 81(4):451–459, doi:10.1016/j.spl.2010.12.003

Holloway ST, Laber EB, Linn KA, Zhang B, Davidian M, Tsiatis AA (2015) DynTxRegime: Methods for Estimating Dynamic Treatment Regimes. URL https://CRAN.R-project.org/package=DynTxRegime, R package version 2.1

Hothorn T, Zeileis A (2015) partykit: A modular toolkit for recursive partytioning in R. Journal of Machine Learning Research 16:3905–3909, URL http://jmlr.org/papers/v16/hothorn15a.html

Hothorn T, Hornik K, Zeileis A (2006) Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics 15(3):651–674, doi:10.1198/106186006X133933

Hubert L, Arabie P (1985) Comparing partitions. Journal of Classification 2(1):193–218, doi:10.1007/BF01908075

Italiano A (2011) Prognostic or predictive? It's time to get back to definitions! Journal of Clinical Oncology 29(35):4718–4718, doi:10.1200/JCO.2011.38.3729

Lang M, Bischl B, Surmann D (2017) batchtools: Tools for R to work on batch systems. The Journal of Open Source Software 2(10), doi:10.21105/joss.00135

Lipkovich I, Dmitrienko A, D'Agostino RB (2016) Tutorial in biostatistics: Data-driven subgroup identification and analysis in clinical trials. Statistics in Medicine doi:10.1002/sim.7064

Loh WY (2002) Regression trees with unbiased variable selection and interaction detection. Statistica Sinica 12(2):361–386, URL http://www.stat.sinica.edu.tw/statistica/oldpdf/a12n21.pdf

Mbogning C, Toussile W (2015) GPLTR: Generalized Partially Linear Tree-Based Regression Model. URL https://CRAN.R-project.org/package=GPLTR, R package version 1.2

Milligan GW, Cooper MC (1986) A study of the comparability of external criteria for hierarchical cluster analysis. Multivariate Behavioral Research 21(4):441–458, doi:10.1207/s15327906mbr2104_5

Seibold H, Zeileis A, Hothorn T (2016) Model-based recursive partitioning for subgroup analyses. International Journal of Biostatistics 12(1):45–63, doi:10.1515/ijb-2015-0032

Seibold H, Hothorn T, Zeileis A (2017) palmtree: Partially additive (generalized) linear model trees. URL https://CRAN.R-project.org/package=palmtree, R package version 0.9-0

Sela RJ, Simonoff JS (2012) RE-EM trees: A data mining approach for longitudinal and clustered data. Machine Learning 86(2):169–207, doi:10.1007/s10994-011-5258-3

Sies A, Van Mechelen I (2017) Comparing four methods for estimating tree-based treatment regimes. The International Journal of Biostatistics Online First, doi:10.1515/ijb-2016-0068

Zeileis A, Hornik K (2007) Generalized M-fluctuation tests for parameter instability. Statistica Neerlandica 61(4):488–508, doi:10.1111/j.1467-9574.2007.00371.x

Zeileis A, Hothorn T, Hornik K (2008) Model-based recursive partitioning. Journal of Computational and Graphical Statistics 17(2):492–514, doi:10.1198/106186008X319331

Zhang B, Tsiatis AA, Davidian M, Zhang M, Laber E (2012) Estimating optimal treatment regimes from a classification perspective. Stat 1(1):103–114, doi:10.1002/sta.411

A. Full factorial simulation

The simulation study described in Section 3 takes a *ceteris paribus* approach and varies one simulation variable at a time while keeping the others at a *standard value*. We did an additional simulation study where we vary all variables, which leads to $8 \cdot 5 \cdot 2 \cdot 4 \cdot 4 \cdot 4 = 5120$ (see Table 1) different scenarios. For each scenario we simulated two data sets and ran all algorithms on each. In the following we show a small selection of interesting graphics based on the simulations. For the full results of the simulation studies we refer to the online material.

Figure 8 shows the marginal results of the ARI for Δ_β, the number of predictive factors, the number of observations and quantitative versus qualitative interactions. We average over the

other simulation variables and the two repetitions. For sake of easy visualisation, we restrict the plotted variable to few levels. Similarly Figures 9 and 10 show the marginal results of the proportion of correct treatment assignment and mean absolute error in estimated treatment effect for the number of predictive factors, Δ_β, the number of observations and quantitative versus qualitative interactions. Figure 11 shows the results for the MAE for $n = 900$ and one prognostic factor to show when LM tree 1 starts to improve (see Section 3.4).

Figure 8 shows that PALM tree can handle simple subgroups with one predictive factor even when the number of observations is low, but the difference in treatment effects must be reasonably high. All other algorithms perform worse, with LM tree 2 and STIMA being the strongest competitors in the low-n-scenarios. OTR performs reasonably well if qualitative subgroups are present. For $n = 500$ the performance of PALM tree rises already at lower levels of Δ_β. The performance of PALM tree and LM tree 2 is very similar and STIMA also performs well. By design OTR ignores any non-qualitative subgroups.

When quantitative treatment subgroups exist, all methods are good at deciding the correct treatment regime (see Figure 9), especially when the number of observations is reasonably high (300). With $n = 100$ PALM tree, LM tree 2, STIMA and even LM tree 1 still perform very well. OTR is the weakest competitor here. With low numbers of observations ($n = 100$), low treatment effect differences ($\Delta_\beta = 0.5$) and qualitative differences, the performance of all algorithms is close to random guessing (0.5), irrespective of the number of predictive factors. With higher Δ_β PALM tree performs reasonably well, followed by LM tree 2, STIMA and OTR (order depending on the number of predictive factors). For $n = 300$ and $\Delta_\beta = 0.5$ STIMA and LM tree 1 perform worst, but STIMA catches up with the other algorithms when $\Delta_\beta = 1.5$, whereas LM tree 1 stays at the bottom. Section 3.3 discusses these results in the context of the results in the star-like simulation study.

Section 3.4 already partly discussed Figures 10 and 11. Figure 10 shows that across different scenarios the MAE increases with increasing number of predictive factors. PALM tree is among the best performers everywhere. In comparison to the other algorithms it performs particularly well in low-n-qualitative scenarios whith $\Delta_\beta = 1.5$.

B. Computation times

The computation times for all methods except STIMA are very reasonable in these applications. For a summary of computation times in the full factorial desing see Table 3. STIMA reached a maximum of 17.4 hours and almost half the models took half an hour or longer.

Table 3: Quantiles of computation times per algorithm in seconds.

	0%	25%	50%	75%	100%
PALM tree	0	0	1	2	7
LM tree 1	0	1	1	2	5
LM tree 2	0	0	1	1	4
OTR	0	0	1	1	2
STIMA	3	233.5	1941	8646.5	62512

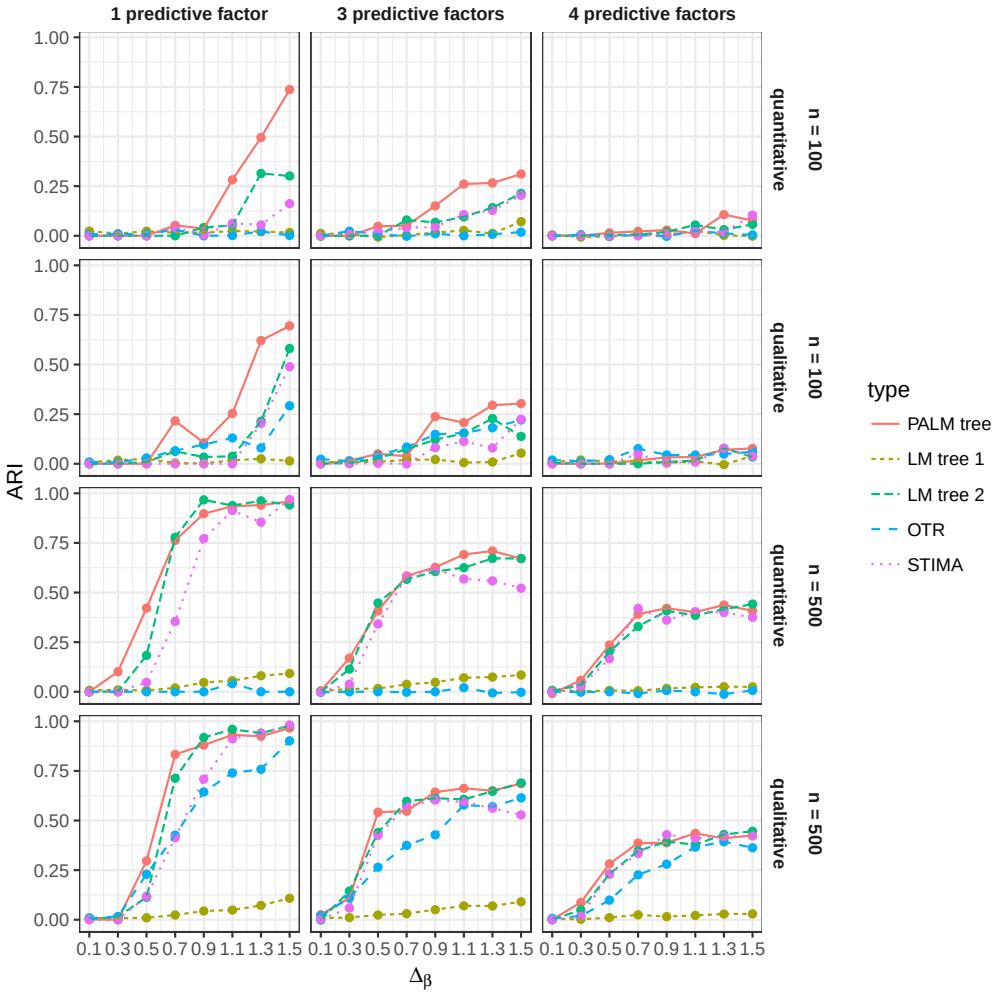


Figure 8: Mean ARI in the full factorial design with two simulated data sets per design (Question 3.1).

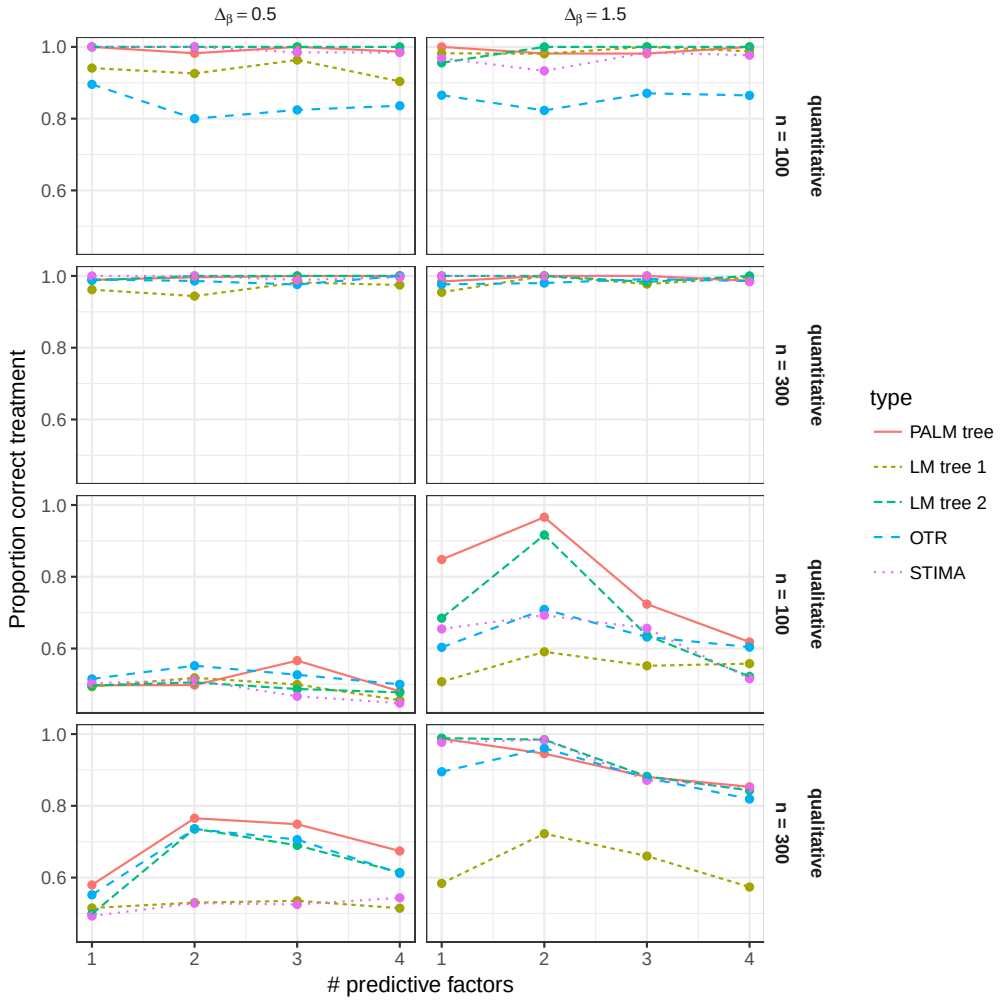


Figure 9: Proportion of observations in all trees where better treatment is correctly identified in the full factorial design with two simulated data sets per design (Question 3.3).

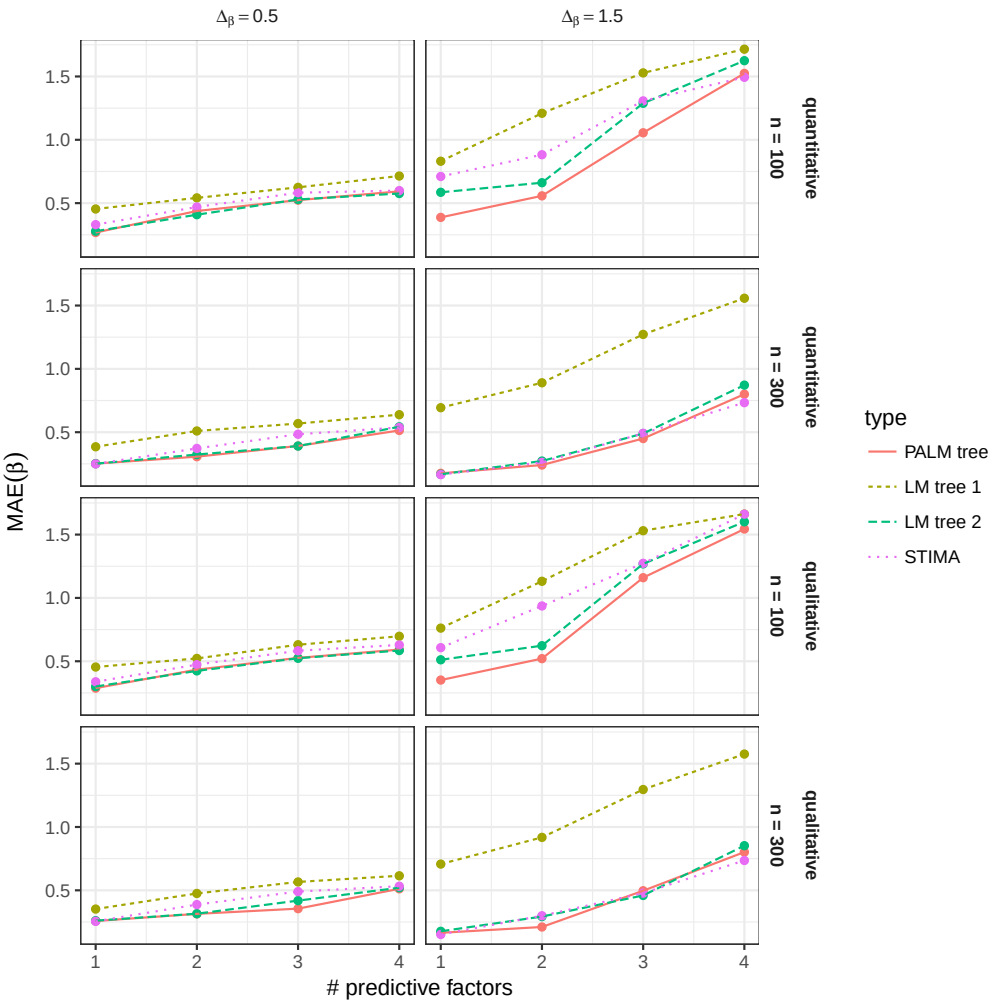


Figure 10: Mean absolute difference between true and estimated treatment effect (mean absolute error, MAE) in the full factorial design with two simulated data sets per design (Question 3.4).

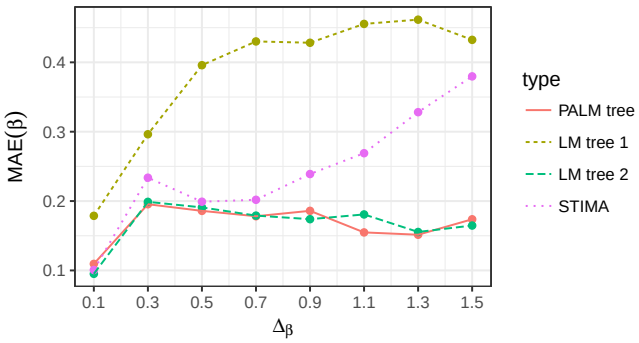


Figure 11: Mean absolute difference between true and estimated treatment effect (mean absolute error, MAE) in the full factorial design with two simulated data sets per design (Question 3.4). Limited data to scenarios with 900 observations and one prognostic factor.

Affiliation:

Heidi Seibold, Torsten Hothorn
Department of Biostatistics
Epidemiology, Biostatistics and Prevention Institute
University of Zurich
Hirschengraben 84
CH-8001 Zurich, Switzerland

Heidi Seibold
Institute for Medical Information Processing, Biometry, and Epidemiology
Ludwig-Maximilans-Universität München
Marchioninistr. 15
81377 Munich

Achim Zeileis
Department of Statistics
Faculty of Economics and Statistics
Universität Innsbruck
Universitätsstr. 15
A-6020 Innsbruck, Austria

PAPER IV

model4you: An R package for personalised treatment effect estimation

Heidi Seibold, Achim Zeileis, Torsten Hothorn

Submitted to the *Journal of Open Research Software*, 2017.

Software paper for submission to the Journal of Open Research Software

(1) Overview

Title
model4you: An R package for personalised treatment effect estimation

Paper Authors
1. Seibold, Heidi; 2. Zeileis, Achim; 3. Hothorn, Torsten

Paper Author Roles and Affiliations
1. PhD student; Biostatistics Department, Epidemiology, Biostatistics & Prevention Institute, University of Zurich
2. Professor; Department of Statistics, Faculty of Economics and Statistics, University of Innsbruck
3. Professor; Biostatistics Department, Epidemiology, Biostatistics & Prevention Institute, University of Zurich

Abstract
Typical models estimating treatment effects assume that the treatment effect is the same for all individuals. Model-based recursive partitioning allows to relax this assumption and to estimate stratified treatment effects (model-based trees) or even personalised treatment effects (model-based forests). With model-based trees one can compute treatment effects for different strata of individuals. The strata are found in a data driven fashion and depend on characteristics of the individuals. Model-based random forests allow for a similarity estimation between individuals in terms of model parameters (e.g. intercept and treatment effect). The similarity measure can then be used to estimate personalised models. The R package *model4you* implements these stratified and personalised models with a focus on ease of use and interpretability so that clinicians and other users can take the model they usually use for the estimation of the average treatment effect and with a few lines of code get a visualisation that is easy to understand and interpret.

Keywords
personalised medicine; subgroup analysis; model-based recursive partitioning; unbiased trees; treatment effect; random forest

Introduction
Studies in various fields randomly assign individuals to one of two groups with different exposure and then measure a response. For example, in clinical trials patients are assigned to one of two treatment groups where usually one treatment group receives a new treatment or drug and the other treatment group receives the standard of care or a placebo. Other examples are in A-B testing in marketing studies or any other two group comparisons such as the mathematics exam discussed below, where students were divided into different exam groups and received slightly

different exam tasks. In the following we will refer to the two groups as *treatment groups* and to the group indicator as *treatment indicator*, which always takes values 0 (individual in first group) and 1 (individual in second group).
Treatment effect estimation is often done using simple models with the binary treatment indicator as only covariate. In the example of a clinical trial the treatment indicator would be 1 if the patient receives the new treatment and 0 if the patient receives standard of care. In R such a simple model can be estimated as follows:

```
base_model <- model(response ~ treatment, data)
```

with response being the response measured, treatment being the treatment indicator and data being the data set containing these variables. The function **model**() can be replaced for example by **lm**() to estimate a linear model, **glm**() to estimate a generalised linear model or **survreg**() to estimate a parametric survival model. These models estimate intercept and treatment effect for all individuals in the data and allow for predicting the response of other individuals given they do or don't receive the treatment of interest.
For cases where the assumption that all individuals have the same intercept and treatment effect is too strict the R package *model4you* offers two options:

1. Model-based trees identify subgroups where within the subgroups the model parameters are similar and between groups the model parameters are different. This is achieved by finding instabilities in the model parameters with respect to a variable (characteristicum) and recursively partitioning the data into groups. If, for example the algorithm finds that men and women have differing treatment effects, the data is partitioned into two subgroups. Details on model-based trees in general can be found in [Zeileis et al., 2008] and for the special use case for stratified treatment effect estimation in [Seibold et al., 2016]. Just a single line of code lets the user compute a model-based tree in R:

```
strat_models <- pmtree(base_model)
```

Note that **pmtree**() uses the data given in the call of the base model. It automatically uses variables not used in the model formula (in the example above response ~ treatment) as potential subgroup defining variables. This can be edited using the zformula argument.

2. Personalised models use model-based random forests to estimate similarity of individuals in terms of model parameters. For each individual a personalised model can be estimated based on a weighted set of the original data, where the similarity measure corresponds to the weight. Details on the personalised models can be found in [Seibold et al., 2017]. Computing personalised models for all observations in the training data is simple:

```
pm_forest <- pmforest(base_model)
pers_models <- pmodel(pm_forest)
```

Again here the potential effect-modifying variables are taken by default as all variable not given in the model formula and can be defined using the zformula argument in **pmforest**().
In the following we will present an example application for model-based trees and personalised models. For this we need to load the package and – to ensure reproducibility – set a random seed. Also for visualisations we need packages *ggplot2*, *ggbeeswarm* and *gridExtra*.

```
library("model4you")
set.seed(2017)

library("ggplot2")
theme_set(theme_classic())
library("ggbeeswarm")
library("gridExtra")
```

Mathematics exam analysis: In 2014 first-year business and economics students at the University of Innsbruck were divided into two examination groups. Group 1 wrote the exam in the morning and group 2 started after the first group finished. The exams for the two groups were slightly different. The data can be accessed and prepared as follows:

```
data("MathExam14W", package = "psychotools")

## scale points achieved to [0, 100] percent
MathExam14W$tests <- 100 * MathExam14W$tests/26
MathExam14W$pcorrect <- 100 * MathExam14W$nsolved/13

## select variables to be used
MathExam <- MathExam14W[ , c("pcorrect", "group", "tests", "study",
                             "attempt", "semester", "gender")]
```

To investigate whether the exam was fair, we asses whether the two groups differ in terms of the percentage of correctly answered exam questions. This can be done using a simple linear model regressing the percentage points of correct anwers on the exam group.

```
bmod_math <- lm(pcorrect ~ group, data = MathExam)
```

The estimtates and confidence intervals of this model can be computed via

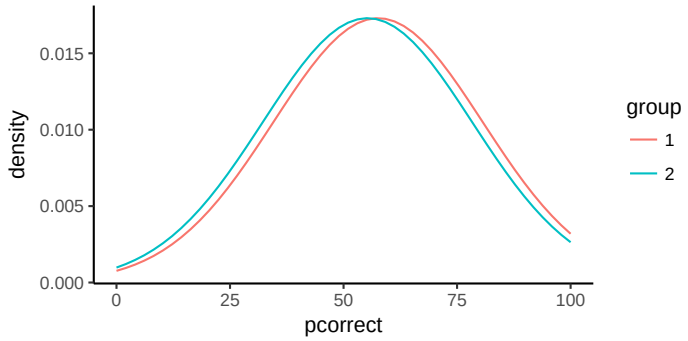


Figure 1: Density estimates of base model for the Mathematics Exam data.

```
cbind(estimate = coef(bmod_math), confint(bmod_math))

##              estimate     2.5 %   97.5 %
## (Intercept) 57.600184 55.122708 60.07766
## group2      -2.332414 -5.698108  1.03328
```

The model can be visualised by plotting the estimated densities (see Figure 1):

```
lm_plot(bmod_math)
```

Both the estimates and confidence intervals and the density curves suggest that there is almost no difference between the two groups. But does this really hold for all types of students?
A tree based on this model can be computed and visualised in only two lines of code:

```
tr_math <- pmtree(bmod_math, control = ctree_control(maxdepth = 2))
plot(tr_math, terminal_panel = node_pmterminal(tr_math,
                                               plotfun = lm_plot))
```

The tree (see Figure 2) divides students based on the percentage of successful online tests. These online tests were conducted biweekly throughout the semester. The largest difference between the two exam groups is in the students who did very well in the online tests (more than 92.3 percent correct). The tree thus gives us much more information on the fairness of the exam than the simple linear model, which is that it does not seem to be fair for students who did very well throughout the semester (at this point we should state that the students self selected into the two exam groups which might also be the reason for differences in exam performance).
Estimating personalised models is almost as simple as the stratified models:

```
forest_math <- pmforest(bmod_math)
pmods_math <- pmodel(forest_math)
```

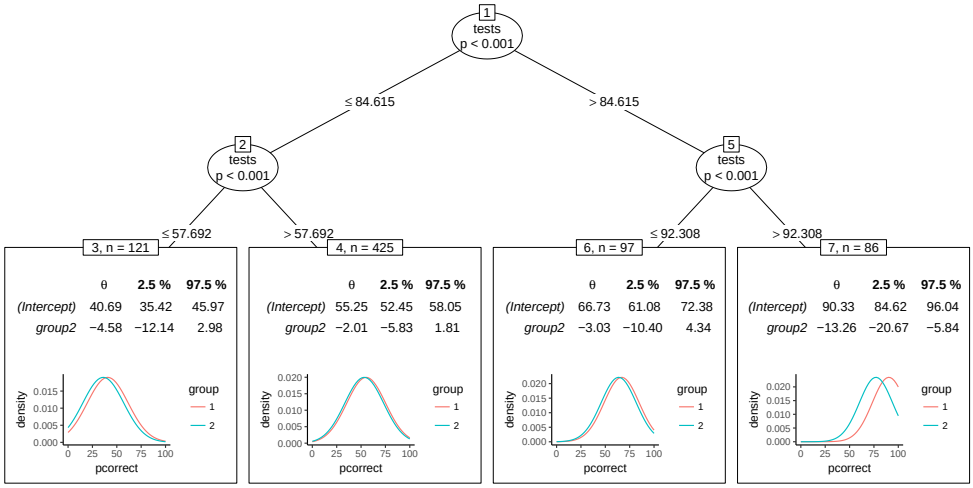
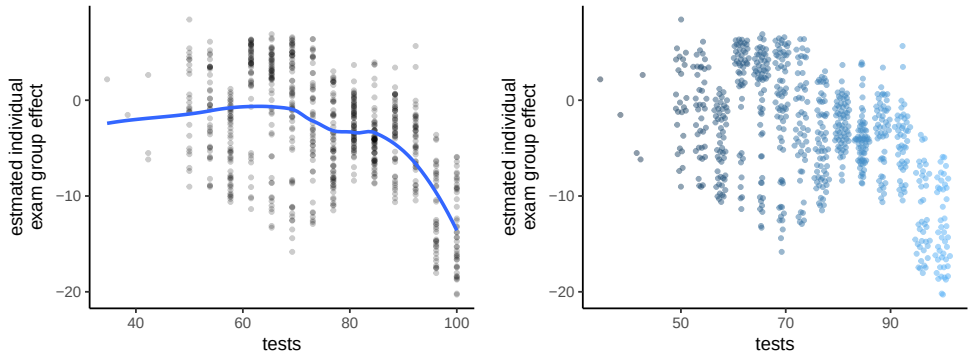


Figure 2: Personalised model tree for the Mathematics Exam datam.

```
## model parameters of first 6 students
head(pmods_math)

##   (Intercept)    group2
## 1    54.80224 -8.449087
## 2    40.58704 -6.119589
## 3    52.56196 -6.776434
## 4    54.58935 -8.898351
## 5    63.88527 -4.960064
## 6    41.29324 -5.991897
```

Dependence plots with the group effect (treatment effect) on the y-axsis and the student characteristics on the x-axis are a good way of visualising the personalised models and for getting knowledge about the interactions between student characteristics and the treatment. Since the percentage of successfull online tests is measured on a grid, a bee plot possibly shows the relationships even better than the scatter plot (both shown in Figure 3).

```
dpdat_math <- cbind(pmods_math, MathExam)

ggplot(dpdat_math, aes(x = tests, y = group2)) +
  geom_point(alpha = 0.2, size = 1) +
  geom_smooth(fill = NA, method = "loess") +
  theme(legend.position = "none") +
  ylab("estmated individual\nexam group effect")

ggplot(dpdat_math, aes(x = tests, y = group2, color = tests)) +
```



(a) Scatter plot.

(b) Bee plot.

Figure 3: Dependence plot for percentage of tests successfully solved.

```
  geom_quasirandom(alpha = 0.5, size = 1) +
  theme(legend.position = "none") +
  ylab("estmated individual\nexam group effect")
```

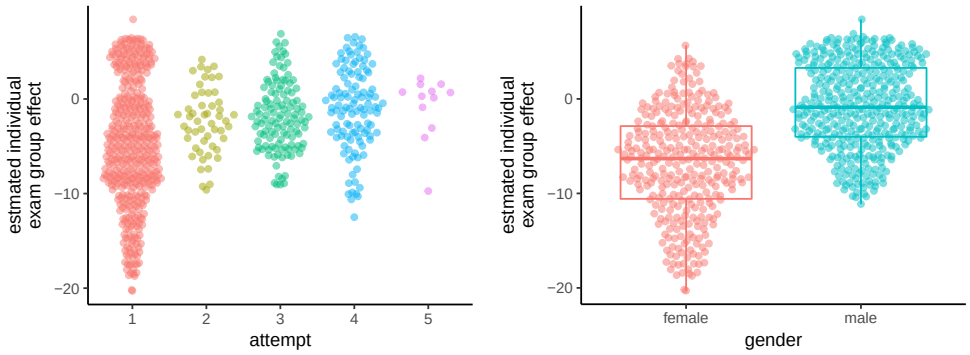
For the number of previous attempts to pass the exam and the gender box plots, bee plots or a combination thereof can be used (Figure 4).

```
ggplot(dpdat_math, aes(x = attempt, y = group2, color = attempt)) +
  geom_quasirandom(alpha = 0.5) +
  theme(legend.position = "none") +
  ylab("estmated individual\nexam group effect")

ggplot(dpdat_math, aes(x = gender, y = group2, color = gender)) +
  geom_boxplot() +
  geom_quasirandom(alpha = 0.5) +
  theme(legend.position = "none") +
  ylab("estmated individual\nexam group effect")
```

With the tools provided by the *model4you* package it is very simple to create understandable stratified and personalised models and compelling visualisations that can be used to communicate theses models.

Implementation and architecture

The R package *model4you* is focused on ease of use and interpretability. Users can take a simple model that they know and understand as basis and simply plug it into **pmtree**() or **pmforest**() depending on whether they want subgroup wise or personalised models. The basis for these functionalities is provided by the *partykit* package which is a widely used R package for trees and forests [Hothorn and Zeileis, 2015, 2017]. The *model4you* package provides wrappers for the well implemented and tested functions partykit::**ctree**() and partykit::**cforest**() and extends the



(a) Number of previous attempts.

(b) Gender.

Figure 4: Dependence plots for the number of previous attempts and gender.

functionalities to allow for the computation of personalised models and to improve usability and interpretability.
The *partykit* package provides the basis for functionalities in other packages namely *glmertree*, *psychotree*, *betareg* (all on CRAN), *trtf*, *disttree*, *lagsarlmtree* and *palmtree* (all available on R-Forge, publishing on CRAN planned).

Quality control
All packages on CRAN undergo standard checks for compatibility with the R package ecosystem. The R package contains examples and tests. These were run and checked on Linux 86_64 and Windows.

(2) Availability

Operating system
Should work on all operating systems that run R.

Programming language
R (version 3.1.0 or higher)

Additional system requirements
None.

Dependencies
R, partykit package (version 1.2 or higher)

List of contributors
Same as the authors: Heidi Seibold, Achim Zeileis and Torsten Hothorn

Software location:
Archive

Name: CRAN
Persistent identifier: `https://cran.r-project.org/package=model4you`

Licence: GPL-2 | GPL-3
Publisher: Heidi Seibold
Version published: The version number of the software archived.
Date published: dd/mm/yy

Code repository

Name: R-forge
Persistent identifier: `https://r-forge.r-project.org/projects/partykit/`
Licence: GPL-2 | GPL-3
Date published: dd/mm/yy

Language
English

(3) Reuse potential
The software is intentionally written to make usage as simple as possible. The most prominent use case are clinical trials where the assumption of an average treatment effect for all patients is too strict and the efficacy of the treatment depends on patient characteristics (e.g. gender, biomarkers, etc.). For subgroup analyses (stratified treatment effects) model-based trees (**pmtree**()) can be used; For personalised treatment effects model-based forests (**pmforest**()) provide a way of estimating similarity between patients and using this similarity measure to estimate personalised models (**pmodel**()). The target audience are people who deal with heterogeneous treatment effects, such as medical researchers, pharmaceutical companies or analysts in marketing (A-B testing). In general the software is useful to researchers dealing with scenarios where two exposures are compared and responses of subjects possibly depend on other variables.
We encourage users to use the party tag on Stackoverflow (`http://stackoverflow.com/questions/tagged/party`) in case of questions or problems.

Acknowledgements
Funding statement
Heidi Seibold and Torsten Hothorn were financially supported by the Swiss National Science Foundation (grants 205321_163456 and IZSEZ0_177091).

Competing interests
The authors declare that they have no competing interests.

References
Torsten Hothorn and Achim Zeileis. partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research*, 16:3905–3909, 2015. URL `http://jmlr.org/papers/v16/hothorn15a.html`.

Torsten Hothorn and Achim Zeileis. *partykit: A Toolkit for Recursive Partytioning*, 2017. URL `http://CRAN.R-project.org/package=partykit`. R package version 1.2-0.

Heidi Seibold, Achim Zeileis, and Torsten Hothorn. Model-based recursive partitioning for subgroup analyses. *International Journal of Biostatistics*, 12(1):45–63, 2016. doi: 10.1515/ijb-2015-0032.

Heidi Seibold, Achim Zeileis, and Torsten Hothorn. Individual treatment effect prediction for amyotrophic lateral sclerosis patients. *Statistical Methods in Medical Research*, 2017. doi: 10.1177/0962280217693034. Online first.

Achim Zeileis, Torsten Hothorn, and Kurt Hornik. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17(2):492–514, 2008. doi: 10.1198/106186008X319331.